Maximum likelihood difference scaling

Laurence T. Maloney Department of Psychology and Center for Neural Science, New York University, New York, NY, USA

Joong Nam Yang NASA Ames Research Center, Moffett Field, CA, USA

We present a stochastic model of suprathreshold perceptual differences based on difference measurement. We develop a maximum likelihood difference scaling (MLDS) method for estimating its parameters and evaluate the reliability and distributional robustness of the fitting method. We also describe a method for testing whether the difference measurement model is appropriate as a description of human judgment of perceptual differences in any specific experimental context.

Keywords: sensory magnitude, proximity, similarity, difference scaling, salience

Introduction

Much research in visual perception involves precise measurement of thresholds. There are well-developed methods for designing experiments to measure threshold (Farell & Pelli, 1998), for analyzing the resulting data (Green & Swets, 1966/1974; Wichmann & Hill, 2001a) and setting confidence intervals on the resulting parameter estimates (Efron & Tibshirani, 1993; Wichmann & Hill, 2001b).

In this article, we develop methods for fitting a model based on difference measurement (Krantz et al., 1971, Chapter 4; See also Roberts, 1979, pp. 134-145) to human judgments of suprathreshold perceptual differences. Researchers have studied suprathreshold color differences (for example, Takasaki, 1966; Ward & Boynton, 19733; Whittle, 1992), contrast differences (McCourt & Blakeslee, 1994), and loudness differences (Schneider, 1980ab), but there have been few studies evaluating appropriate experimental methodology and fitting procedures. There is only one series of papers (Schneider, 1980ab; Schneider et al., 1974) reporting a substantial application of difference measurement to a psychophysical problem. We will describe this work in Conclusions and compare it with our own.

We describe the difference measurement model in detail in a later section. The simplest way to explain it is to describe the kind of task that it is intended to model. Consider the upper and lower pair of color samples shown in Figure 1. For most observers the two colors samples in each pair are readily distinguishable[1], as evidenced by an appearance of a sharp border separating them. The differences between both pairs are suprathreshold, but at the same time most observers, forced to choose the pair where the difference is color is 'greater', would pick the upper pair.

The configuration in Figure 1 is an example of a typical trial from a difference scaling experiment. The observer is asked to examine two pairs of stimuli ('a quadruple') and to select the pair with the larger perceptual difference. However, there is more to the experimental design than judgment of quadruples. All four of the color samples in Figure 1 fall on a line in color space, illustrated in Figure 2. The upper pair of samples in Figure 1 are copies of the 1st and 5th color samples in Figure 2. The lower pair are copies of the 7th and 8th. Over the course of an experiment, the observer is asked to make this judgment for a large number of quadruples of color samples, all drawn from Figure 2.

Our goal is to assign numbers $\psi_1, \psi_2, \cdots, \psi_{11}$ to the 11 stimuli in Figure 2 so that the absolute differences between scale values accurately predicts the observer's judgments. If, for example, the observer always judges the difference of the upper pair in Figure 2 to be greater that that of the lower, then, whatever the choice of the numbers $\psi_1, \psi_2, \cdots, \psi_{11}$, they should satisfy

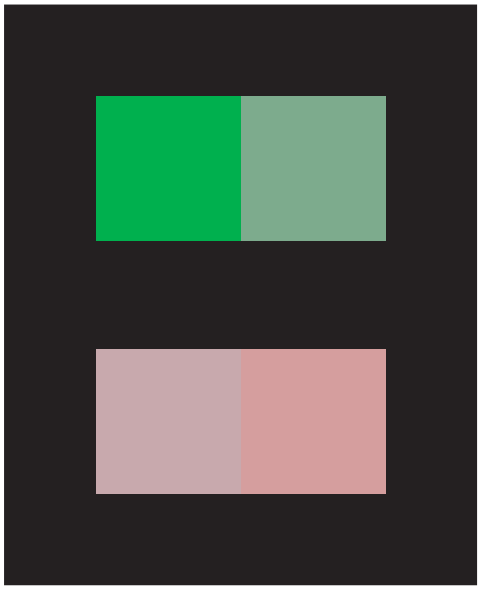

Figure 1. A difference scaling quadruplet. The observer's task is to examine the upper pair of color patches and the lower pair, and to select whichever pair has a larger difference.

DOI 10:1167/3.8.5 Received February 25, 2003; published October 7, 2003 ISSN 1534-7362

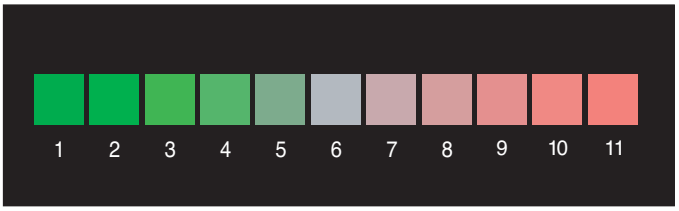

Figure 2. Color patches that fall on a line in LMS space. Eleven color patches that fall on a straight line in LMS space (Wyszecki & Stiles, 1982, pp. 119ff). The goal of difference scaling is to assign 11 numbers to these stimuli so that differences in scale values predict perceived suprathreshold color differences.

$|\psi_2 - \psi_5| > |\psi_7 - \psi_8|$. The Maximum Likelihood Difference Scaling (MLDS) method we develop takes as input the 2-alternative forced choice (2AFC) data for a few hundred quadruples. It assigns scale values that best predict the experimental data. The MLDS method is the natural, suprathreshold counterpart to standard methods for estimating threshold differences (most recently the elegant papers by Wichmann & Hillis, 2001ab). In Figure 3, we plot estimates of ψ_i for the stimuli in Figure 2, for one naive observer. The estimates were obtained using the MLDS method. They are based on 330 2AFC difference scaling trials, each similar to the trial illustrated in Figure 1.

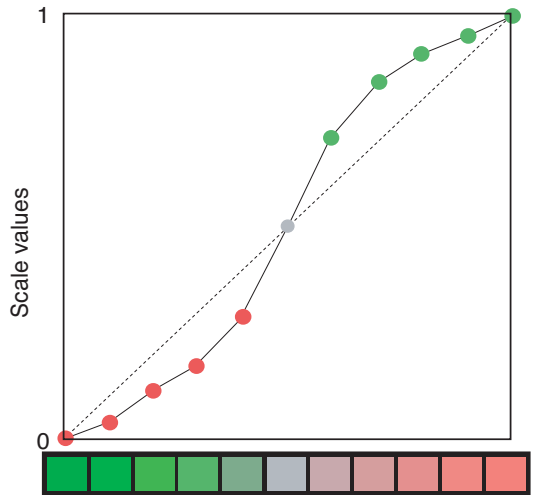

Figure 3. An estimated difference scale. The stimuli are reproduced along the horizontal axis. The values plotted along the vertical axis are scale values assigned to 11 colored patches by a maximum likelihood difference scaling procedure described later in the article. The evident, sigmoidal shape of the interpolated curve indicates 'crispening': the perceived differences of pairs of color patches are exaggerated near the neutral point.

The resulting function is sigmoidal, and roughly symmetric about the neutral color patch. The decrease in slope with increasing saturation in both the 'green' and the 'red' directions is consistent with the well-known 'crispening' effect (Takasaki, 1966; Whittle, 1992). These error bars are computed by an application of a resampling method (Efron & Tibshirani, 1993), also described here.

In Figure 4, we give a second example of an application of difference scaling to the measurement of the effect of applying a particular image compression algorithm (vector quantization, described in Gersho & Gray, 1991) on 'image quality'. The degree of compression is controlled by an arbitrary, univariate parameter, γ, that ranges from 1 ('no compression') to 30 ('maximum compression'). Compression reduces the size of the image by a factor $1/\gamma$. Figure 4 contains examples of three images with different degrees of compression as shown. The reader will likely agree that the difference between $\gamma = 1$ ('no compression') and $\gamma = 15$ is less than the difference between $\gamma = 15$ and $\gamma = 24$. The evident aliasing artifacts in the last image make it difficult to interpret the contents of the scene portrayed.

Figure 4. An image at three levels of compression. The compression algorithm is described in detail in Knoblauch et al. (1998). It has one free parameter, γ, that controls degree of compression. The compressed file takes up $1/\gamma$ as much space as the original file. Images corresponding to γ values of 9, 15, and 24 are shown.

Knoblauch et al. (1998) applied MLDS to derive a scale of perceived distortion in the image as a function of compression. The results of applying a MLDS procedure based on 210 trials are shown in Figure 5. The horizontal axis is the degree of image compression, γ. The vertical axis is the estimated scale values. It is evident that the image compression algorithm has little effect on the image when $\gamma \leq 15$ but, above that point, small changes in γ result in larger perceived differences in the image. Of course, the results of the difference scaling do not tell us whether the observer prefers the uncompressed image ($\gamma = 1$) to the most compressed image ($\gamma = 24$). It does suggest that so long as $\gamma \leq 15$, the benefits of image compression come with relatively little change in perceived image quality.

We will not pursue these examples further. The key issue we wish to emphasize here is that we could only accept these results if we trust the experimental and

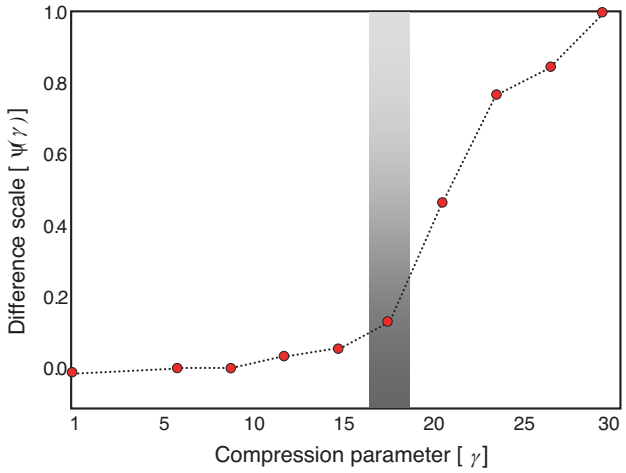

Figure 5. An estimated difference scale. The nominal value plotted on the horizontal axis is γ. The values plotted along the vertical axis are scale values assigned to the 10 compressions of the same image by a maximum likelihood difference scaling procedure described later in the article.

computational methods that we have employed. In the next section, we describe a standard difference scaling experiment and develop a stochastic difference scaling model that we will use to predict the probability that an observer will pick a given pair of stimuli over a second given pair. We show how to fit this model to experimental data and, in following sections we investigate its reliability. We show that we can use standard resampling methods to compute confidence intervals for estimated scale values. We test whether failures in the distributional assumptions underlying the method affect the results of the fitting procedures (i.e., whether the estimation method is robust).

We chose the two examples to illustrate the wide range of problems to which we can apply difference scaling. We also note that scaling methods are controversial. It is very likely that the reader, in considering the second example above ('image quality'), was led to doubt that there is any unidimensional perceptual variable corresponding to image quality. He or she might be less suspicious concerning the color example since there is a considerable literature on comparison of color differences.

One of the most serious failings of the scaling literature is that often there is no evident way to reject a model of proximity judgments as simply inappropriate as a model for human perception of differences (but see Hutchinson & Tversky, 1986; Gerrig, Maloney & Tversky, 1991). In a later section of the paper, we return to this point and propose methods for testing whether the difference scaling model is an adequate model for what the observer has done in a particular experiment.

A Stochastic Model of Difference Measurement

The experimenter selects N stimuli, $S_1,\cdots,S_N$. Associated with each stimulus S_i is a real number, ϕ_i, and the stimuli are numbered so that $\phi_1 < \phi_2 < \cdots < \phi_N$. In our two examples, the values ϕ_i corresponded to degree of image compression, and distance from the origin in a color space, respectively. The experimenter presents an observer with *quadruples*, $(S_i, S_j; S_k, S_l)$, and asks him to judge which pair S_i,S_j or S_k,S_l is more salient.

It will prove convenient to replace $(S_i, S_j; S_k, S_l)$ by the simpler notation $(i, j; k, l)$. In traditional difference measurement, the experimenter wishes to determine *scale values* $\psi_1,\cdots, \psi_N$ corresponding to the stimuli, $S_1,\cdots,S_N$, such that, given a quadruple, $(i, j; k, l)$, the observer judges S_i, S_j to be further apart than S_k, S_l precisely when,

$$| \psi_j - \psi_i | > | \psi_l - \psi_k | \ . \tag{1}$$

We will usually order the stimuli in a quadruple so that $\psi_j > \psi_i$, and $\psi_l > \psi_k$ and then we can omit the absolute value signs in Equation 1.

It is unlikely that human observers will be so reliable in judgment as to satisfy the criterion just given, particularly if the differences $\psi_j - \psi_i$ are near 0. Accordingly, we develop a model that allows the observer to exhibit some stochastic variation in judgment. Let $l_{ij} = \psi_j - \psi_i$, the *length* of the interval S_i, S_j. Then we assume that the *decision variable*[2] employed by the observer is

$$D(i, j;k, l) = l_{ij} - l_{kl} + \varepsilon, \tag{2}$$

where ε is a Gaussian random variable with mean zero and standard deviation $\sigma > 0$. Given the quadruple, $(i, j; k, l)$, the observer is assumed to select the pair S_i, S_j precisely when,

$$D(i, j;k, l) > 0 \tag{3}$$

The proposed model is an equal-variance Gaussian signal detection model (Green & Swets, 1966/1974) where the signal is the difference in the lengths of the intervals, $l_{ij} - l_{kl}$. When this length is small, negative or positive, relative to σ, we expect the observer, presented with the same stimuli, to give different, apparently inconsistent judgments.

The choice of a Gaussian random variable is arbitrary and the assumption that the error term is additive and independent of the lengths of the intervals under judgment can also be questioned. In a later section, we will evaluate the distributional stability of the estimation procedure that we describe next.

Maximum Likelihood Difference Scaling

At first glance, it would appear that the stochastic difference scaling model just presented has $N+1$ free parameters $\psi_1,\cdots, \psi_N$ together with the standard deviation of the error term, σ. However, any linear transformation of the $\psi_1,\cdots, \psi_N$ together with a corresponding scaling of σ results in a set of parameters that predicts exactly the same performance as the original parameters. Without any loss of generality, we can set $\psi_1 = 0$ and $\psi_N = 1$, leaving us with the $N-1$ free parameters, $\psi_2,\cdots, \psi_{N-1}$ and σ. We describe next how to estimate these parameters by the method of maximum likelihood. The fitting procedure is similar to that of maximum likelihood estimates used in fitting psychometric functions (Wichmann & Hill, 2001a).

Suppose that, on a particular trial, the observer sees the quadruple $(i, j; k, l)$ and judges the first interval to be larger. Given any choice of the free parameters $\psi_2,\cdots, \psi_{N-1}$ and σ, we can compute the probability that this will occur. It is simply the probability that the decision variable is positive for this choice of free parameters:

$$\mathrm{P}[D(i, j;k, l) > 0 \mid \psi_2,\cdots, \psi_{N-1}, \sigma] \quad . \tag{4}$$

This is simply $\Phi_\sigma(l_{kl} - l_{ij})$ where

$$\Phi_\sigma(x) = P[\varepsilon \le x] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{u^2}{2\sigma^2}} du \tag{5}$$

is the cumulative distribution function of the Gaussian random variable ε. If the observer instead selects the second interval, then the probability that he does so is just $1 - \Phi_\sigma(l_{kl} - l_{ij})$. If we code the choice of the first interval as $R = 1$ and the choice of the second interval as $R = 0$, then the probability of the subject's response is written as,

$$P(r \mid \psi_2,\cdots, \psi_{N-1}, \sigma) = \Phi_\sigma(\psi_j - \psi_i - \psi_l + \psi_k)^R \ \mathrm{X}$$

$$(1 - \Phi_\sigma(\psi_j - \psi_i - \psi_l + \psi_k))^{(1-R)} \tag{6}$$

where we have expanded the term $l_{ij} - l_{kl}$ in terms of the parameters $\psi_2,\cdots, \psi_{N-1}$ so that it is evident that the probability on the left hand side does depend upon the (unknown) values of the free parameters.

Of course, there is little to be learned from a single trial. We wish to compute the probability of a pattern of responses, R_t, across many trials, numbered $t = 1,\cdots, T$. Let the quadruple on trial t be $(i_t, j_t; k_t, l_t)$, and define the difference in lengths on trial t to be $\Delta_t = \psi_j - \psi_i - \psi_l + \psi_k$. Then the probability of the observed pattern of responses for any choice of the free parameters $\psi_2,\cdots, \psi_{N-1}$, and σ, is,

$$P[R_1,\cdots,R_T \mid \psi_2,\cdots,\psi_{N-1},\sigma] = \prod_{t=1}^{T} \Phi_\sigma(\Delta_t)^{R_t} [1 - \Phi_\sigma(\Delta_t)]^{-R_t} \tag{7}$$

which is also the likelihood $\mathrm{L}[\psi_2,\cdots, \psi_{N-1}, \sigma \mid R_1, \cdots,R_T]$ of any particular choice of the free parameters given the (known) outcome of the experiment $R_1, \cdots,R_T$. We form maximum likelihood estimates of the free parameters, $\hat{\psi}_2,\cdots,\hat{\psi}_{N-1}$ and $\hat{\sigma}$, by finding the values of these parameters that maximize the equation above. There is no closed form for such a solution but it is very easy to compute the estimates by standard numerical optimization methods (Hanselman & Littlefield, 2001, Chapter 22; Press et al., 1992, Chapter 10).

Maximum likelihood estimates have several desirable asymptotic properties. As we base the estimates on more and more data, any bias in the estimates vanishes (Mood, Graybill & Boes, 1974). Further, as the amount of data increases, the maximum likelihood estimates have a variance that converges to the minimum possible for an unbiased estimator (Mood et al., 1974, pp. 358-371). These are desirable properties, and, consequently, most modern estimation procedures are based on maximum likelihood or its close Bayesian relatives (Berger, 1985). However, it is important to assess whether the estimators are biased when used with the amount of data collected in typical experiments. The results based on difference scaling in Figures 3 and 5 are intriguing, but it is certainly important to check whether they are contaminated by biases in estimation. Further, it is important to determine how variable estimates are for any specified number of trials, to estimate confidence intervals for the parameters $\psi_2,\cdots, \psi_{N-1}$. Before turning to an examination of the 'small sample' bias and variability of the MLDS estimates just described, we need to discuss some of the practical aspects of designing and carrying out a difference scaling experiment.

Designing a Difference Scaling Experiment

The experimenter selects N stimuli, $S_1,\cdots,S_N$ and prepares a list of quadruples, $(i, j; k, l)$. It is convenient to avoid repetitions in the list of indices. If, for example, the observer is presented with the quadruple $(1,2;1,3)$ then he can potentially note that the 'right' answer is signaled by the ordering of the second stimulus in each interval. He can then base his judgment solely on the basis of an ordering of the stimuli. The total possible number of stimuli with distinct indices $1 \le i < j < k < l \le N$ is $\binom{N}{4}$, which can be written out as,

$$D(N) = \frac{1}{24}\left[\begin{array}{l}(N-4)^4 + 10(N-4)^3 \\ +35(N-4)^2 + 50(N-4) + 24\end{array}\right] \tag{8}$$

Some values are tabulated for small values of N in Table 1. Given any quadruple, (i, j, k, l), with $1 \le i < j < k < l \le N$, it may be presented to the subject in eight different ways. For example, in the color stimuli in Figure 2, we could present (i, j) above and (k, l) below or vice versa. Wherever we chose to present (i, j), we can present S_i on the left and S_k on the right, or vice versa. Similarly for (k, l). Since the ordering of the stimuli $(i < j)$ should be obvious to the observer, there is little point in bothering to randomize left-right and we do not. We do, however, randomize the locations of (i, j) and (k, l), effectively flipping a coin to decide which one goes above or below, or, for temporal forced-choice, which one goes first or second.

For ten stimuli, then, one first pass through all possible intervals in random order requires 210 judgments (Table 1). Assuming that these forced-choice judgments take no more than a 5 seconds each, it is possible to go through 210 judgments in about 20 minutes or less.

Table 1. The number of quadruples in a complete design on a specified number of stimuli.

Number of Stimuli	Number of Quadruples
10	210
12	495
15	1365
20	4845

These values are computed using Equation 8.

For 20 or more stimuli, the total number of possible trials becomes too large to contemplate. However, it is still possible to carry out difference scaling and parameter estimation using only a fraction of the possible trials. In a later section, we investigate how many trials are actually needed to establish a difference scale for N stimuli for values of N larger than 10.

Of course, we can also repeat the series of 210 trials for 10 stimuli as many times as we like. In the next section we examine how the bias and reliability of the estimates vary with the number of trials.

Evaluating Bias and Variability

Bias and Variablity as a Function of Number of Trials

The maximum likelihood fitting procedure just described is asymptotically unbiased and asymptotically minimum variance (Mood et al., 1974). However, for experiments involving realistic numbers of trials and typical experimental protocols it is important to assess the bias and standard deviation for experiments involving realistic numbers of trials and choice of quadruples. We evaluated both by Monte Carlo computation.[3]

For each experiment, we specified a function $\psi(s) = s^\gamma$ relating the nominal scale values to the difference scale values, a value of the standard deviation σ of the error term in the comparison process, and n levels of the nominal scale value, $S_1, \cdots, S_N$. We then simulated the difference scaling observer in one or more repetitions of a complete design on n stimuli. The simulated data were fit to the model to obtain estimates $\hat{\psi}(s_i)$ and an estimate of $\hat{\sigma}$ of σ. We repeated this cycle of simulation and fitting 1000 times and computed the mean and standard deviations of the estimates $\hat{\psi}(s_i)$ and $\hat{\sigma}$. The mean is an estimate of the expected value $E[\hat{\psi}(s_i)]$ of the estimator $\hat{\psi}(s_i)$ and the standard deviation, an estimate of the standard deviation $SD[\hat{\psi}(s_i)]$ of the estimator.

In Figure 6 we plot the results of such Monte Carlo replications for five different choices of the power function $\psi(s) = s^\gamma$ with $n = 10$ levels of the nominal scale value. A complete design for $n = 10$ includes 210 quadruples. We ran experiments in which the simulated observer completed one repetition of a complete design (210 trials) and also four repetitions of a complete design (840 trials). The value of σ was set to 0.2, a value typical of the estimated variability of our subjects in the color experiments.

The power functions $\psi(s) = s^\gamma$ are plotted as curves and the estimates of $E[\hat{\psi}(s_i)]$ are plotted as points that, if the estimator were unbiased, should lie precisely on the corresponding curve. For 210 trials, the estimator is slightly biased for the case $= 2$ for nominal values near zero (the points are slightly above the curve). For 840 trials, this bias is no longer visible. Overall the maximum likelihood estimation procedure returns estimates that are close to unbiased, even with a modest number of trials.

The estimates of the standard deviations $E[\hat{\psi}(s_i)]$ are plotted as error bars attached to each point. The error bars allow us to predict how close to the true difference scale values $\psi(s_i)$ the estimates could be expected to fall in any given experiment. The simulated experimental conditions closely mimic the conditions of the difference scaling of color in the first example and indicate that the experimental protocol and fitting methods are not introducing any important bias, at least when the distributional assumptions are satisfied.

The Effect of Random Subsampling of the Set of Possible Trials.

With 15 or 20 stimuli, the number of trials (quadruples) in a complete design is large (Equation 8). We simulated experiments with $N = 10$, 12, 15, or 20 stimuli in which observers completed only a subset of a complete design, either 10% or 50% of the quadruples chosen at random. The mapping from nominal values to scale values was always the power function $\psi(s) = s^{1/2}$. For example, with 20 stimuli, the number of quadruples

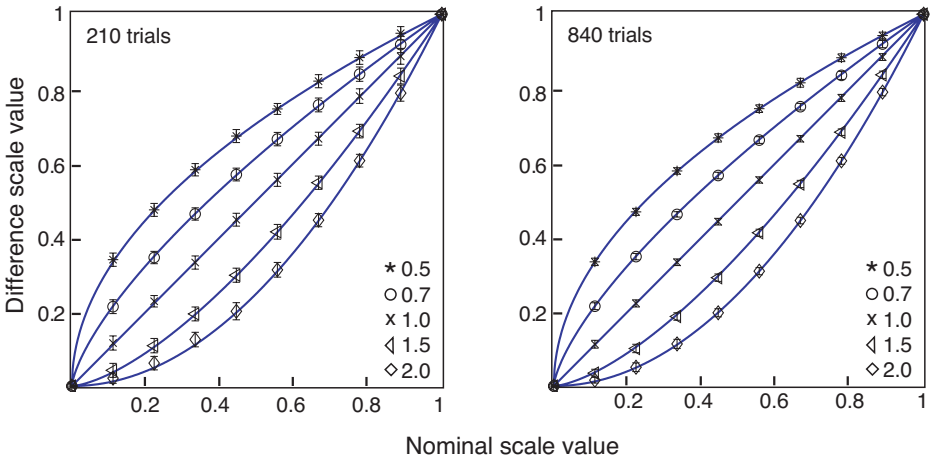

Figure 6. Bias and standard deviation in MLDS estimation. The plots are summaries of the mean and standard deviations for estimates of the 10 scale parameters of a simulated observer. The σ value for the simulated observer was set to 0.2, a typical value for human observers in color scaling experiments such as the one described in the introduction. The nominal scale values are equispaced in the interval [0,1]. They are plotted on the horizontal axis and the average of 200 simulated estimates on the vertical. The true difference scale was a power function ψ^{γ} of the nominal scale values. We range five series of simulations with γ set to 0.5, 0.7, 1.0, 1.5, and 2, respectively. The power functions are plotted as smooth curves. To the extent that the points fall on the curves shown, the estimates are unbiased. The error bars are ±1 SD for a single estimate obtained in repeated simulations. This is the uncertainty in estimation associated with performing a difference scaling experiment on single observer and fitting the resulting data by MLDS.

(Table 1) is 4845. In the 10% condition, the observer would complete 484 of these quadruples, chosen at random. In the 50% condition, he or she would complete 2426 quadruples. The responses on this subset of quadruples would be the basis for computing the likelihood function that is then maximized to obtain estimates of the difference scale values. For comparison, we also simulated experiments in which 100% of the possible quadruples were used. We refer to this condition as '100% subsampling'. There were a total of 12 experimental conditions (four choices of N, three choices of subsampling).

We repeated each experimental condition 200 times and computed the standard deviation of the scale value estimates for each scale value, and averaged these standard deviation estimates across scale values. In Figure 7, we show a log-log plot of mean standard deviation versus the number of quadruples used, i.e. we plot the mean standard deviation for 10% subsampling of the 4845 possible trials for 20 stimuli versus the number of trials, 484. The subsampling rates are coded by color, the number of stimuli by shape. The results for 10% subsampling of the 210 possible trials for 10 stimuli (21 trials) did not always converge[4] and is omitted from the plot. There were no convergence failures for larger rates of subsampling or larger numbers of stimuli (and correspondingly larger numbers of possible trials).

The variance of estimates decreased as the subsampling rate increased and as the total number of possible trials increased. It is evidently possible to obtain reliable estimates with as many as 20 stimuli with 10% subsampling. We have approximately the same variance for 484 trials taken as a 10% subsampling of the 4845 possible trials for 20 stimuli and for 495 trials taken as a 100% sample of the 495 possible trials for 12 stimuli (Figure 7). The logarithm of variance (the square of the

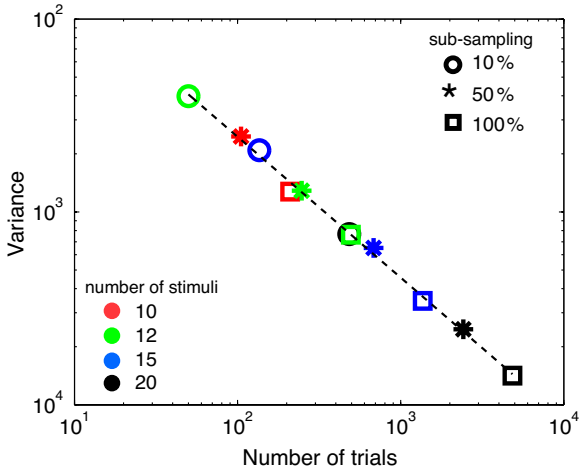

Figure 7. The effect of randomly subsampling the possible trials for N stimuli. The number of possible quadruples grows rapidly with the number of stimuli (Equation 8). We simulate difference scaling experiments in which only a fraction (10%, 50%, 100%) of the possible quadruples are used. We did so for $N = 10, 12, 15, 20$ stimuli. The σ value for the simulated observer was set to 0.2. The mean standard deviation of the estimates of the $N-2$ scale values that are free to vary is plotted versus the number of trials. It is approximately a linear function with slope -0.73.

standard deviation) is roughly a linear function of the logarithm of the total number of trials over the range we considered. We estimated the slope of the best-fitting line by linear regression. This line, with slope -0.73, is shown in the plot.

These results indicate that it is feasible to use difference scaling with 20 or more stimuli. If M is the total number of trials used in estimate a difference scale by means of MLDS, the expected variance of the parameter estimates is roughly proportional to $M^{3/4}$ over the range considered (replacing 0.73 by 3/4). That is, the expected standard deviations of the parameter estimates (the error bars in Figure 6) are roughly proportional to $M^{3/8}$.

Distributional Robustness

Effect of Changes of Distributional Family

In developing the maximum likelihood fitting method, we modeled the observer's judgment of a quadruple $(i, j; k, l)$ as comparison of a stochastic decision variable $D(i, j; k, l)$ to a threshold. The decision variable was Gaussian and its standard deviation did not depend on the magnitudes of the stimuli or the differences of stimuli under comparison. In this section, we examine how violations of either of these assumptions affect the estimates returned by the fitting method.

We first developed models of observers who carried out the difference scaling task by forming a decision variable based on the difference of the lengths of two intervals, perturbed by additive random error, just as before:

$$D(i, j\,;\, k, l) = l_{ij} - l_{kl} + \varepsilon\ . \tag{9}$$

The additive error was independent and identically-distributed, but it was not always Gaussian in form. We simulated observers where the distribution was Uniform on the interval $(-s_u, s_u)$, Laplacian or Cauchy in form. The Laplacian has the probability density function,

$$f(x) = \frac{1}{2s_l} e^{-\left|\frac{x}{s_l}\right|} \tag{10}$$

and the Cauchy has probability density function,

$$f(x) = \frac{1}{\pi s_c}\left(1 + \left(\frac{x}{s_c}\right)^2\right)^{-1}\ . \tag{11}$$

All three distributions are symmetric about 0. We simulated replications of a difference scaling experiment with each of the Uniform, Cauchy, and Laplacian Observers in turn, and fitted the resulting data with the Gaussian maximum likelihood methods described previously. We adjusted the scale parameters s_u, s_c, s_l, so that the fitted estimate $\hat{\sigma}$ of observer variability was roughly the same as that observed in our experiments with human observers. The Laplacian and Cauchy are 'high-tailed' distributions that are often used in testing the distributional robustness of fitting methods based on the Gaussian distribution. They tend to have relatively large numbers of 'outliers' compared to the Gaussian. If the results of the maximum likelihood fitting method change markedly when the true distribution of the observer is not Gaussian, then the fitting method is not distributionally robust. Conversely, if we get the same estimates of the scale values ψ_i despite changes in the error distribution, then the Gaussian assumption can be treated as a convenience, not a crucial assumption we employ in using the method.

In Figure 8, we plot the mean fitted values $\hat{\psi}_i$ obtained using the MLDS procedure versus true values of ψ_i for the Uniform, Cauchy and Laplacian observers. As can be seen, changes in distributional form lead to only small biases. The MLDS procedure is distributionally robust.

Effect of Non-Uniform Variance

For all of the simulated observers we assumed that the variability of the additive error was independent of the magnitude of the difference between the interval lengths, an homogeneity of variance assumption. In Figure 9, we report the results of simulating a Non-Uniform Variance Observer and fitting the resulting data assuming homogeneity of variance. The variance was proportional to the absolute value of the difference $l_{kl} - l_{ij}$ and the constant of proportionality was chosen so that the fitted value $\hat{\sigma}$ was roughly the fitted variability of human observers in our experiments. In Figure 9, we plot the mean fitted values $\hat{\psi}_i$ obtained using the Gaussian MLE procedure versus true values of ψ_i for the Non-Uniform Variance observers (compare Figures 6 and 9). In summary, the Gaussian MLE procedure described above is distributionally robust over the range of conditions considered.

Axiomatic Issues and Validation

Interpretation of the results of any psychophysical estimation procedure depends on the assumptions made in modeling and fitting the data. In estimating a sensory threshold we assume, for example, that the threshold we are estimating is roughly stable over the period of time employed in measuring it. In considering any particular application of difference scaling, the experimenter can seek to determine whether the model used in fitting the data was appropriate to the observer and the particular experimental situation. We have just seen that the results of the fitting procedure are not much affected by failures of the Gaussian distributional assumption. However, we

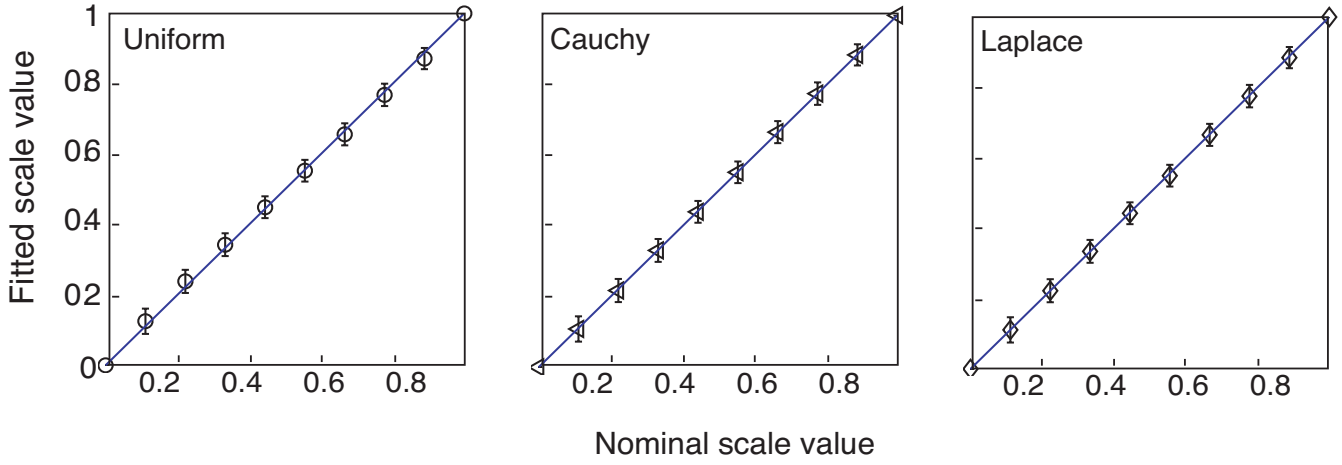

Figure 8. Distributional robustness: other distributions. A simulated observer whose underlying additive random error is drawn from one distributional family is fitted by an experimenter who assumes that the observer's underlying additive error is Gaussian. The true scale values of the observer are plotted on the horizontal axis, and the average of 200 estimated scale values on the vertical. To the extent that the resulting plot falls on the 45 degree line, the fitting procedure is distributionally robust for the particular distributions selected. The three curves correspond to uniform error, Cauchy error, and double exponential error (see text). The total number of simulated trials was 210.

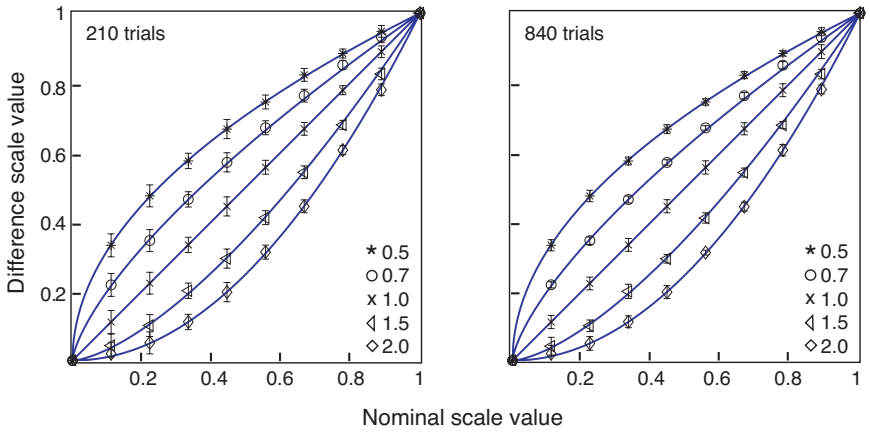

Figure 9. Distributional robustness: non-homogeneous variance. A simulated observer whose underlying additive random error is Gaussian but with standard deviation increasing with the magnitudes of the intervals compared is fitted by an experimenter who assumes that the observer's underlying additive error is Gaussian with constant standard deviation. The nominal values of the observer are plotted on the horizontal axis, the average of 200 estimated scale values on the vertical. To the extent that the plotted points fall on the corresponding curves, the MLDS fitting procedure is not affected by nonhomogeneous variance.

can consider other basic assumptions made in applying difference scaling methods as we have described.

Krantz et al. (1971) contains derivations of necessary and sufficient conditions that an observer must satisfy if we are to conclude that his judgments can be described by a difference scaling model. Two of these represent testable claims about human performance.

The first is the Ordering Property. The observer must be able to reliably order the stimuli $S_1, \cdots, S_N$ in agreement with the ordering of the scale values $\psi_1 < \psi_2 < \cdots \psi_N$. Often, in difference scaling applications, this ordering property is evidently satisfied. Ordering the color samples in Figure 2 is not difficult for anyone with normal color vision.

The second is the Six-Point Property, illustrated in Figure 10. There are two groups of three stimuli whose indices are i, j, k and i', j', k', respectively. For convenience, let us assume that $i<j<k$ and $i'<j'<k'$ as indicated in the figure. Otherwise, we put no restrictions on the ordering of the six stimuli. Suppose that a non-stochastic observer considers the quadruple (i, j, i, j') and judges that $ij < i'j'$. On some other trial, he considers (j, k, j', k') and judges that $(jk<j'k')$. Now, given the quadruple, (i, k, i, k'), there is only one possible response consistent with the difference scaling model. He must

choose $ik < i'k'$. The reasoning behind this constraint is illustrated in the figure and it can be demonstrated directly from the model. Since the observer has judged that $ij < i'j'$, we known that

$$\psi_j - \psi_i > \psi_{j'} - \psi_{i'} \quad (12)$$

and since he has judged that $jk < j'k'$, we also know that

$$\psi_k - \psi_j > \psi_{k'} - \psi_{j'} \ . \quad (13)$$

Adding, Equations 12 and 13, we have,

$$\psi_k - \psi_i > \psi_{k'} - \psi_{i'} \quad (14)$$

and the observer must judge that $ik<i'k'$, as stated above. For the non-stochastic observer, even one violation of this six-point condition would allow us to conclude that there was no consistent assignment of scale values $\psi_1, \psi_2, \cdots \psi_N$. in a difference scaling model that could predict his or her judgments in a difference scaling task.

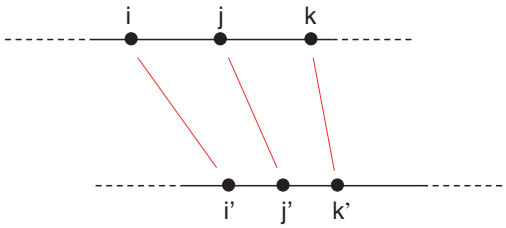

Figure 10. The six-point condition. If the observer's judgments do not satisfy the six-point condition illustrated here, then there is no difference scale that captures these judgments. Suppose that the observer judges the interval (i, j) greater than the interval (i, j') and the interval (j, k) greater than the interval (j', k'). Then he must judge the interval (i, k) greater than the interval (i', k'). Of course, the observer may occasionally violate a six-point condition because of stochastic variability in his judgments. See the text for a reformulation of the six-point condition in a form that allows for stochastic variability in response.

Of course, the non-stochastic observer is an implausible model of human behavior: human judgments in difference scaling tasks are not deterministic. In MLDS, decisions are based on a decision variable $D(i, j; k, l)$ and, given any six points i, j, k, i', j', k' there is a non-zero probability that the stochastic observer will violate the six-point condition. In particular, suppose that $\psi_j - \psi_i$ is only slightly greater than $\psi_{j'} - \psi_{i'}$, $\psi_k - \psi_j$ is only slightly greater than $\psi_{k'} - \psi_{j'}$, and that $\psi_k - \psi_i$ is only slightly greater than $\psi_{k'} - \psi_{i'}$. Then we might expect that the observer has roughly a probability of 0.5 of judging $ij > i'j$ and similarly with the other two quadruples. Hence, he has an appreciable change of judging $ij > i'j', jk > j'k'$ and $ik < i'k'$, violating the six-point property.

In deciding whether an observer's judgments are consistent with the six-point property and the difference scaling model, we must take into account that the stochastic observer is expected to occasionally violate the six-point condition. We need to develop a method for deciding whether a particular pattern of violations of the six-point condition is sufficiently suspicious to warrant rejecting the difference scaling model as a model of the subject's performance. First, we enumerate all possible sextuples i, j, k, i', j', k' and denote them $\psi_1, \cdots, \psi_{N6}$. Given ML estimates of the scale values used by the human observer $\hat{\psi}_1, \hat{\psi}_2 \cdots, \hat{\psi}_N$, and the observer's standard deviation, $\hat{\sigma}$, we can compute the probability that the observer would violate any given six-point condition $\psi_a = (i, j, k, i', j', k')$ by judging $ij > i'j'$, $jk > j'k'$ but $ik < i'k'$. This probability is denoted p_a[5].

We can now count the number of times, V_a, that the observer has violated the six-point condition τ_a during the course of the experiment and number of times he has satisfied it, S_a. The likelihood of this outcome, for this one choice of a six-point condition is,

$$\Lambda_6^a = p_a^{V_a} (1 - p_a)^{S_a} \ , \quad (15)$$

and we can now compute the overall likelihood of the observed outcome by taking the products of the likelihoods of all possible six-point conditions:

$$\Lambda_6 = \prod_{a=1}^{N_6} \Lambda_6^a \quad (16)$$

Of course, if the observer never judges one of the three pairs of outcomes in a sextuple, perhaps because of subsampling, then $V = S = 0$ and the corresponding likelihood is one.

We have developed a Monte Carlo method for comparing the observed likelihood of the observer's six-point judgments to the expected distribution of this likelihood for the ideal stochastic observer with scale values that are the same as the fitted scale values $\hat{\psi}_1, \hat{\psi}_2 \cdots, \hat{\psi}_N$ and standard deviation $\hat{\sigma}$ for the observer.

In Figure 11, we show the histogram of the logarithm of replications of the six-point likelihood Λ_6 for a simulated ideal stochastic observer whose parameters were the fitted parameters of Observer SHC in Figure 3. The logarithm of the actual six-point likelihood Λ_6 for the observer is marked by a vertical line. The values in the histogram are measures of the extent to which a simulated observer, perfectly described by the MLDS model, could reasonably be *expected* to violate the six-point condition over the course of a particular experiment.

If the human observer SHC does in fact satisfy the MLDS model, then his or her Λ_6 should tend to fall near the center of this distribution. Had the actual value fallen

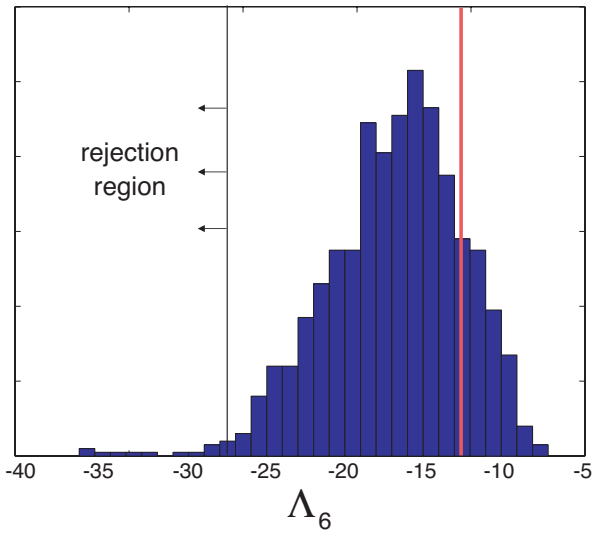

Figure 11. Testing the six-point condition. The plot is a histogram of the logs of estimated likelihoods of six-point violations Λ_6 for 1000 simulations of a single observer. The simulations were based on the fitted parameter values of the observer SHC whose results are plotted in Figure 3. The 5th percentile of the distribution is marked by a vertical black line. We would reject the six-point hypothesis if the likelihood of the observer's six-point violations fell to the left of this line (too improbable). The likelihood of the six-point violations for this observer is plotted as a red line. It falls well above the 5th percentile mark, suggesting that the observer's pattern of six-point violations is consistent with the difference scaling model.

below the α percentile of the histogram, we would instead reject the hypothesis that the difference scale model is appropriate for this observer at the α–level. The simulated MLDS observer has such an extreme value of Λ_6 on fewer than α proportion of the replications. In that case, we judge that the pattern of six-point failures in the observer's data is too improbable and, following the ordinary logic of hypothesis testing, we reject the hypothesis that the MLDS model is appropriate for SHC at the α–level.

In Figure 11, for example, an estimate of the $\alpha = 0.05$ level is marked by a vertical black line. This estimate is simply the 5th percentile of the histogrammed values. The shaded rejection region is to the left of the line. The observer's true value does not fall within the rejection region and we do not reject.

Conclusions

Difference scaling provides a rigorous means for assessing suprathreshold differences and it is the natural complement to methods for measuring thresholds along sensory continua. By means of examples, we illustrated the range of its applicability. We have presented a stochastic model of judgments in tasks involving ranking of suprathreshold perceived difference of stimuli that fall on a one-dimensional continuum. We described methods for estimating the parameters of the model from data and described how to design difference scaling experiments. We found that the estimator is nearly unbiased in realistic experimental designs and that it is distributionally robust. While the number of possible trials grows rapidly with the number of stimuli, we found that the experimenter could obtain reliable estimates with only a small fraction of all of the possible comparisons.

In many respects, MLDS is a stochastic, unidimensional form of metric and non-metric multidimensional scaling methods (Shepard, 1962ab, 1974; Torgerson, 1958). In MLDS the error model that explains observers' inconsistencies in judging the same stimuli is made explicit and plays a central role. The use of maximum likelihood methods guarantees that MLDS is asymptotically unbiased and minimum variance. The simulations reported here suggest that MLDS exhibits little bias and narrow confidence intervals for parameters with only modest amounts of data. In that respect, it resembles psychometric fitting methods (Wichmann & Hill, 2001ab). There are no comparable guarantees for traditional MDS methods.

We also showed how to test necessary and sufficient conditions for difference measurement in the case where observers' estimates are perturbed by stochastic error. If both the ordinal and six-point conditions hold, then observers' performance is consistent with a difference measurement model. If an observer's performance is not consistent with the difference scaling model, then one of these two conditions must fail. By translating the tests of these conditions into statistical hypothesis tests, we, of course, leave open the possibility that the difference measurement model will fail to be rejected because the corresponding statistical test lacks power. However, nothing better is to be expected if we do not wish to reject the model because of minor stochastic variation in an observer's judgment. There are analogous tests of necessary conditions for multidimensional representations (Hutchinson & Tversky, 1986; Gerrig et al., 1991) but, currently, no tests of necessary and sufficient conditions.

The methods and analyses presented in this article complement earlier work by Schneider and colleagues (Schneider, 1980ab; Schneider et al., 1974). The fitting method proposed by Schneider and colleagues is non-metric and does not explicitly model stochastic variability in observers' responses. Consequently they cannot directly address hypothesis testing or issues of bias and efficiency. MLDS is asymptotically unbiased and efficient (minimum variance) because it estimates parameters by maximizing likelihood. There is no comparable guarantee for the method of Schneider and colleagues. Further, MLDS lends itself to hypothesis testing, most notably in evaluating the match between human performance and the axioms of difference measurement. Schneider and

colleagues can argue that the number of axiom violations in a data set is small, but cannot reasonably decide whether the violations observed constitute grounds for rejecting the difference measurement model. As we noted above, for certain configurations where there are many near ties in interval lengths, we might expect a large number of six-point violations from an observer whose behavior is actually consistent with the six-point condition. MLDS allows us to determine whether the pattern of six-point violations observed is inconsistent with the six-point condition.

There are other, classical methods for assigning scale values to stimuli on a one-dimensional continuum (McIver & Camines, 1981) of which Thurstonian scaling is perhaps the best known (Thurstone, 1927). These methods differ markedly from MLDS. They rely heavily on confusions between adjacent stimuli on a scale[6], while MLDS compares both large and small intervals. Scales estimated using Thurstonian scaling, for example, are known to be sensitive to the choice of the error distribution. We note that there is no reason to expect that a scale based on measurements of just-noticeable-differences between stimuli could be used to predict supra-threshold perceptual differences.

The basic task in MLDS is a comparison of intervals (a quadruple). Other scaling methods use other tasks, notably the method of triads or duo-triads (Torgerson, 1958, pp. 262ff; Frijters, 1979; Ennis & Mullen, 1986). In one form of the method of triads, the observer considers three stimuli, A, B, and C and decides whether the difference between A and B is greater than, or less than, the difference between A and C. We can reinterpret this judgment as a comparison of intervals: the observer is deciding whether the interval AB is greater than, or less than, the interval AC. The method of triads, in this form, is evidently a special case of the quadruples task in which the four stimuli need not be distinct. In its earliest form (Torgerson, 1958, pp. 262ff), the observer must choose which of the three intervals, AB, AC, BC is smallest (or alternatively, largest). Again, the method of triads reduces to a comparison of intervals. Torgerson describes other methods that amount to comparison of intervals. Any method that permits the experimenter to establish an interval scale on a one-dimensional continuum of stimuli must, directly or indirectly, involve comparison of intervals.

The number of possible triads for N stimuli is evidently less than the number of possible quadruples. For ten stimuli, for example, there are 135 triads and 210 quadruples composed of four distinct stimuli. It seems, at first glance, that an experiment involving triads should involve fewer trials than an experiment involving quadruples. However, we showed in the 'Evaluating Bias and Variability' Section that the number of *actual* trials, not the number of *possible* trials, determines the standard deviations of the resulting scale estimates. The experimenter, who seeks a lower standard deviation in estimating a scale on 10 stimuli, is advised to repeat the 210 possible trials as many times as needed. In our simulations we show the effect of repeating the 210 possible trials up to four times (840 trials total). Conversely, we show that the experimenter who wishes to scale a large number of stimuli can use only a subset of the possible quadruples and get reliable results. There is no reason to use one subset of possible quadruples (e.g. the 'triadic quadruples') rather than another simply because the former subset has fewer elements.

In all of our analyses we considered only quadruples on four distinct stimuli with each quadruple used in a simulation chosen from the set of such quadruples with equal probability. This choice was arbitrary. It would be of interest to determine whether experiments based on some subsets of quadruples are more efficient than others. It would also be of interest to develop adaptive procedures that select the next quadruple to be presented based on the observer's previous responses.

Scaling methods can be controversial (Gescheider, 1988). In advancing MLDS, we view it as a method for fitting a plausible model (difference measurement) to data. The model contains a precise description of how the observer arrives at each judgment including possible stochastic variation in response. The fitting methods and methods for setting confidence intervals and testing hypotheses are based on current practice in statistics. Moreover, MLDS is remarkably distributionally robust. Consequently, we suggest that the only controversy that might be relevant is whether the difference measurement model is an adequate model of human judgment of suprathreshold perceptual differences, a controversy that is best resolved by further experimentation with the method.

Acknowledgments

This research was funded in part by Grant EY08266 from the National Institute of Health. LTM was also supported by grant RG0109/1999-B from the Human Frontiers Science Program. We thank Kenneth Knoblauch and Michael Landy for comments on an earlier draft of this article. We especially thank Kenneth Knoblauch and colleagues for permission to use the images and data in Figures 4 and 5. Commercial relationships: none.

Footnotes

[1] The color illustrations here may not reproduce accurately when printed or displayed on uncalibrated monitors. The descriptions in the text are correct when the images are accurately displayed.

[2] We can also formulate a multiplicative model in which the observer compares ratios of ratios of magnitude rather than differences of differences and where the error is

multiplicative. The resulting model, with an appropriate choice of multiplicative error distribution, is behaviorally indistinguishable from the additive error form described here. The multiplicative model becomes the additive if we simply take the logarithm of the decision variable.

[3] All of the computer code used to fit data and to perform Monte Carlo simulations is available from the authors. It is written in the C language.

[4] When the experiment consists of a small number of quadruples, there may not be enough data to determine a solution. That is, many different choices of scale values share the same maximal likelihood.

[5] At first glance, it may appear that we have neglected to consider a second possible six-point violation involving these points: $ij < i'j'$, $jk < j'k'$ but $ik > i'k'$. This second pattern of violation is, however, counted when we consider the sextuple.

[6] Unlike most variants of Thurstonian scaling that are based on analysis of a measured confusion matrix, Watson & Kreslake (2001) propose a method for finding just-noticeable-differences (JNDs) within a given physical scale.

References

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. New York: Springer.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman Hall.

Ennis, D. M. & Mullen, K. (1986). A multivariate model for discrimination methods. *Journal of Mathematical Psychology*, *30*, 206-219.

Farell, B. and Pelli, D. G. (1998). Psychophysical methods, or how to measure a threshold and why. In J. G. Robson and R. H. S. Carpenter (Eds.), *A Practical Guide to Vision Research*. New York: Oxford University Press.

Frijters, J. E. R. (1979). Variations of triangular method and the relationship of its unidimensional probabilistic models to three-alternative forced-choice signal detection theory models. *British Journal of Mathematical and Statistical Psychology*, *32*, 229-241.

Gerrig, R. J., Maloney, L. T. & Tversky, A. (1991), Validating the dimensional structure of psychological spaces: Applications to personality and emotions. In D. R. Brown & J. E. K. Smith [Eds.], *Frontiers of Mathematical Psychology*. New York: Springer-Verlag. pp. 138-165.

Gersho, A. & Gray, R. M. (1991), *Vector Quantization and Signal Compression*. Kluwer Academic Publishing.

Gescheider, G. A. (1988), Psychophysical scaling. *Annual Review of Psychology*, *39*, 169-200. [PubMed]

Green, D. M. & Swets, J. A. (1966/1974) *Signal Detection Theory and Psychophysics*. New York: Wiley. Reprinted 1974, New York: Krieger.

Hanselman, D & Littlefield, B. (2001), *Mastering MATLAB 6; A Comprehensive Tutorial and Reference*. Upper Saddle River, NJ: Prentice-Hall.

Hutchinson, J. W. & Tversky, A. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*, 3-22.

Knoblauch, K., Charrier, C., Cherifi, H., Yang, J. N. & Maloney, L. T. (1998) Difference scaling of image quality in compression-degraded images [Abstract]. Perception (Supplement), *27*, 174.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1): Additive and Polynomial Representation, Chapter 4. New York: Academic Press.

McCourt, M. E. & Blakeslee, B. (1994). Contrast-matching analysis of grating induction and suprathreshold contrast perception. *Journal of the Optical Society of America A*, *11*(1), 14-24. [PubMed]

McIver, J. P. & Camines, E. G. (1981). *Unidimensional scaling*. Beverly Hills, CA: Sage Publications.

Mood, A., Graybill, F. A., & Boes, D. C. (1974). Introduction to the theory of statistics, 3rd Edition, pp. 358-371. New York: McGraw-Hill.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*, 2nd Ed. Cambridge, UK: Cambridge University Press.

Roberts, F. S. (1979). Measurement theory with applications to decision making, utility and the social sciences. Reading, MA: Addison-Wesley.

Schneider, B. (1980a). Individual loudness functions determined from direct comparisons of loudness intervals. *Perception and Psychophysics*, *28*(6), 493-503. [PubMed]

Schneider, B. (1980b). A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika*, *45*(3), 357-372.

Schneider, B. Parker, S. & Stein, D. (1974). The measurement of loudness using direct comparisons of sensory intervals. *Journal of Mathematical Psychology*, *11*, 259-273.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*, 125-140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*, 219-246

Shepard, R. N. (1974), Representations of structure in similarity data: Problems and prospects. *Psychometrika*, *39*, 373-421.

Takasaki, H. (1966). Lightness change of grays induced by change in reflectance of gray background. *Journal of the Optical Society of America*, *56*, 504-509. [PubMed]

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Ward, F. & Boynton, R. M. (1973). Scaling of large chromatic differences. *Vision Research*, *14*, 943-949.

Watson, A. B. & Kreslake, L. (2001). Measurement of visual impairment scales for digital video. Proceedings of the SPIE, *4299*, 79-89. [Article]

Whittle, P. (1992). Brightness, discriminability and the "Crispening Effect." *Vision Research, 32*, 1493-1507. [PubMed]

Wichmann, F. A. & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling and goodness of fit. *Perception & Psychophysics*, *63*, 1293-1313. [PubMed]

Wichmann, F. A. & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*, 1314-1329. [PubMed]

Wyszecki, G & Stiles, W. S. (1982). Color science: Concepts and methods, quantitative, data and formulas, 2nd Ed. New York: Wiley.