ANNUAL REVIEWS

Annual Review of Vision Science

Measuring and Modeling Visual Appearance

Laurence T. Maloney[1,2] and Kenneth Knoblauch[3,4]

[1]Department of Psychology, New York University, New York, New York 10003, USA; email: laurence.maloney@nyu.edu

[2]Institut des Etudes Avancées de Paris, 94004 Paris, France

[3]Université Lyon, Université Claude Bernard Lyon 1, INSERM, Stem Cell and Brain Research Institute U1208, 69500 Bron, France; email: ken.knoblauch@inserm.fr

[4]National Centre for Optics, Vision and Eye Care, Faculty of Health and Social Sciences, University of South-Eastern Norway, 3616 Kongsberg, Norway

Annu. Rev. Vis. Sci. 2020. 6:519–37

First published as a Review in Advance on May 18, 2020

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-030320-041152

Keywords

scaling, suprathreshold, maximum likelihood difference scaling, maximum likelihood conjoint measurement, MLDS, MLCM, proximity, diagnostics

Abstract

In studying visual perception, we seek to develop models of processing that accurately predict perceptual judgments. Much of this work is focused on judgments of discrimination, and there is a large literature concerning models of visual discrimination. There are, however, non-threshold visual judgments, such as judgments of the magnitude of differences between visual stimuli, that provide a means to bridge the gap between threshold and appearance. We describe two such models of suprathreshold judgments, maximum likelihood difference scaling and maximum likelihood conjoint measurement, and review recent literature that has exploited them.

1. INTRODUCTION

Albedo: the proportion of incident light reflected from the surface; its subjective correlate is defined as lightness

Since Fechner's (1860) initial development of psychophysics, researchers have used measures of sensory discrimination in developing and testing models of visual perception. The trichromatic theory of color vision is based on characterization of the lights that can and cannot be discriminated by the typical observer (Kaiser & Boynton 1996). Much research concerning spatial vision (for example, Pelli & Bex 2013) and perception of motion (for example, Nishida et al. 2018) is based on discrimination measures.

Discrimination measures are readily linked to maximum likelihood and Bayesian ideal observer models (Geisler 1989, Green & Swets 1966, Maloney & Zhang 2010). Ernst & Banks (2002), for example, showed that the discriminability of individual shape cues approximately controlled the weight given to each cue in combination, as predicted by statistical models of optimal cue combination (Landy et al. 1995). There are well-developed experimental methods for assigning numerical estimates of discriminability (Green & Swets 1966, Maloney & Zhang 2010), for analyzing the resulting data (Green & Swets 1966, Wichmann & Hill 2001), and for setting confidence intervals on the resulting parameter estimates (Efron & Tibshirani 1994, Knoblauch & Maloney 2012, Wichmann & Hill 2001). The combination of a visual judgment (discrimination) and a model of the judgment process [a linking hypothesis in Brindley's (1970) terms] has proven to be immensely fruitful in vision research.

In this review, we consider other judgments and models that plausibly capture the observer's perception of suprathreshold appearance: how red or glossy or viscous or rough stimuli appear to be (Fleming et al. 2015). Researchers have studied suprathreshold differences with a variety of methods (for example, McCourt & Blakeslee 1994, Takasaki 1966, Ward & Boynton 1974, Whittle 1992). The methods that we describe provide systematic approaches to measuring and modeling suprathreshold appearance by means of simple, forced-choice, perceptual judgments.

The stimuli in **Figure 1*a*** differ in albedo, increasing in albedo from left to right. We denote the albedo of the *j*th stimulus, left to right, by φ_j, and by design $\varphi_1 < \varphi_2, \ldots, < \varphi_N$ with $N = 7$. The stimuli increase in albedo left to right, and most of the pairs of stimuli are readily discriminable to the typical observer. Thus, measurements of discrimination would be minimally informative. However, given any pair of distinct stimuli $[\varphi_i, \varphi_j]$ with $\varphi_i < \varphi_j$, we can consider the apparent difference between the two stimuli. We call such a pair an interval and sometimes abbreviate notation to *ij* for convenience. (We use the Greek letter φ to denote physical quantities such as albedo and the Greek letter ψ for the internal representations of these qualities in the models considered below.)

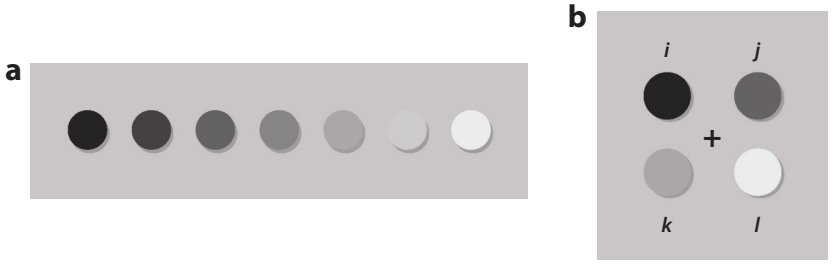

Figure 1

(*a*) The circles are identical except for albedo (the proportion of light reflected), denoted, from left to right, $\varphi_1, \ldots, \varphi_N$. (*b*) A quadruple judgment. Is the difference between the upper pair *ij* greater than or less than the difference between the lower pair *kl*? The plus sign is a fixation target. The letters do not appear in the display.

Intuitively, we are trying to capture a judgment of the relative magnitude of intervals, and a typical such judgment is schematized in **Figure 1*b***. The two intervals *ij* and *kl* correspond to pairs of stimuli taken from the stimuli in **Figure 1*a***, and the judgment is straightforward: We ask the observer to judge which interval is greater or, equivalently, which difference is greater. While $i < j$ and $k < l$ by convention, the order of the four indices i, j, k, and l is not otherwise constrained. The intervals could be nonoverlapping (13 versus 56) or overlapping (14 versus 25). One interval could even be contained in the other (15 versus 24). The intervals are typically large and readily discriminable.

For the example in **Figure 1*b***, the typical observer will select the lower pair *kl* as larger. The pair of intervals in **Figure 1*b*** form a quadruple, and judging which interval is greater is a quadruple judgment. The experimental designs that we use, based on quadruple judgments, are referred to as the method of quadruples.

If there are N stimuli, then there are $N_Q = N(N-1)/2$ possible intervals. We can ask observers to order not the stimuli but rather the N_Q intervals (pairs of stimuli). As N increases, the number of possible pairs of intervals N_Q increases quadratically, far more rapidly than the number of stimuli N. If N is 10, then the number of pairs of pairs is 990. In theory, we could ask an observer to judge all possible quadruples (pairs of intervals) for N stimuli. In practice, we can confine attention to a remarkably smaller subset of interval comparisons (see below; for a discussion, see Maloney & Yang 2003).

In Section 2, we introduce maximum likelihood difference scaling (MLDS) (Maloney & Yang 2003), a model of interval judgment. Section 2.1 concerns the model, Section 2.2 describes an application in detail, and Section 2.3 briefly reviews experimental applications of the method. In Section 3, we describe a different but closely related method, maximum likelihood conjoint measurement (MLCM), based on a different judgment and model.

In previous work, Knoblauch & Maloney (2008, 2012) and Maloney & Yang (2003) described the mathematical framework of both MLDS and MLCM in detail and advised on how to set up experiments and analyze data. In this review, we emphasize the outcome of recent applications of the methods by ourselves and others.

Our current implementations of the scaling methods are in the statistical language R (Knoblauch & Maloney 2008, Knoblauch et al. 2019, R Core Team 2019). By embedding the methods in a statistical framework, we inherit diagnostic tests and also model selection methods (Anderson & Burnham 2004, Knoblauch & Maloney 2012). A version has also been implemented in MatLab (Kingdom & Prins 2016), and another in Python (**https://github.com/computational-psychology/mlds**) provides a wrapper to our R package.

The roots of MLDS and MLCM can be traced to earlier work in mathematical psychology, notably the multivolume *Foundations of Measurement* series (Krantz et al. 1971, 1989; Roberts 1985; Suppes et al. 1990). A marriage between this rich and powerful literature and modern statistical methods has led to new methods that provide powerful ways to model human perception.

2. MAXIMUM LIKELIHOOD DIFFERENCE SCALING

2.1. The Model

We begin with N stimuli indexed along a physical scale, such as the scale in **Figure 1*a*** and the quadruple judgment illustrated in **Figure 1*b***. In most applications, N is typically chosen to be 10 or greater: The stability and accuracy of MLDS estimates for lower values have not been systematically explored. We denote the physical scale values as $\varphi_1 < \varphi_2 \cdots < \varphi_N$. The goal of MLDS is to find a difference scale $\psi_1 < \psi_2 < \cdots < \psi_N$ that can predict the observer's ordering

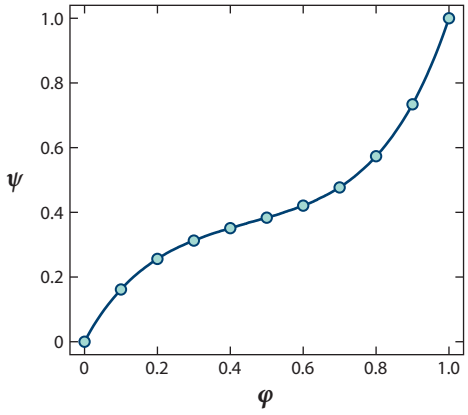

Figure 2

A hypothetical difference scale with ψ_i plotted versus φ_i for $i = 1, \ldots, N$.

of the $N(N-1)/2$ intervals ij. Define the unsigned perceptual length of an interval ij as $\delta_{ij} = |\psi_i - \psi_j|$, and let $ij \prec kl$ denote the case in which the observer judges interval ij to be less than interval kl. We wish to find a difference scale such that

$$ij \prec kl \text{ if and only if } \delta_{ij} < \delta_{kl}. \qquad 1.$$

If we succeed at assigning scale values that predict perceived ordering of intervals, then it seems that we have captured $N(N-1)/2$ ordering judgments in only N parameters. It is actually fewer, since we can add a constant to the values $\psi_1 < \psi_2 < \cdots < \psi_N$ without changing any of the predicted orderings. We may equally multiply them by a positive constant without changing the predicted orderings. We can, for example, choose the parameterization $\psi_1 = 0$ and $\psi_N = 1$ without any loss of generality (Knoblauch & Maloney 2012, Maloney & Yang 2003).

2.1.1. The nonstochastic model. For simplicity, we first describe the model as if observers' judgments were nonstochastic: Presented with the same stimuli, the observer would always make the same judgment. We then modify it to take into account the noisiness of the observer's judgments. The nonstochastic model is just a didactic tool and plays no role in MLDS.

In **Figure 2**, we plot ψ_i versus φ_i, a graphical representation of a difference scale that superficially resembles a psychophysical function (for examples of relating the psychometric and psychophysical functions, see Hillis & Brainard 2005, 2007a,b). The vertical axis in this case is used to predict perceived differences, δ_{ij}, in appearance. In addition to the plotted points we add a hypothetical continuous curve. The physical stimuli are often drawn from a continuum, and the resulting difference scale is intended to generalize beyond the physical sample. The implementation of MLDS by Knoblauch & Maloney (2012) also allows fitting of functions drawn from continuous families specified by the user. Such curves, specified by only a few parameters, can be compared with the points using model comparison methods. We can compare a parametric model fit to a less restrictive, nonparametric model fit. If the physical scale is continuous, then we can interpolate and invert the curve to estimate a scale with uniform perceptual spacing (see, for example, Rogers et al. 2016).

2.1.2. The stochastic model. The nonstochastic model fails if observers are not consistent in their judgments, as is often the case in experiments. Suppose that the experimenter repeatedly

presents two intervals *ij* and *kl*. The observer judges $ij \prec kl$ on some presentations but $kl \prec ij$ on others. Evidently, we cannot assign values ψ_j consistent with the observer's contradictory judgments. Intuitively, we might expect such inconsistencies when the perceived length of *ij* and the perceived length of *kl* are very close to each other.

The approach taken in MLDS is to include a stochastic component in the judgment process (judgment noise) itself. If the experimenter uses the packages available (for example, the R package available at **https://cran.r-project.org/web/packages/MLDS/index.html**), then he need not be concerned with the details of the actual MLDS fitting procedure. Nevertheless, we briefly sketch the model.

We can rewrite Equation 1, adding a random Gaussian variable ϵ with mean 0 and variance σ^2, as

$$ij \prec kl \text{ if and only if } \delta_{kl} - \delta_{ij} + \epsilon > 0. \qquad 2.$$

Any choice of $\psi_2 < \psi_3 < \cdots < \psi_{N-1}$ and variance σ^2 induces probabilities of each of the ordering judgments $ij \prec kl$ for all i, j, k, l (recall that ψ_1 and ψ_N are not free parameters, but instead are set to fixed values, for example, 0 and 1, respectively). We can compute the likelihood (probability) of any pattern of an observer's independent responses and then vary the parameters to maximize this likelihood. The result is maximum likelihood estimates of the $N-1$ parameters $\psi_2 < \psi_3 < \cdots < \psi_{N-1}$ and σ^2. The judgment noise parameter σ^2 is a measure of the observer's inconsistency in judgment.

Other choices of parameterization are possible, however, and may be desirable in particular circumstances. In **Figure 2**, we normalize the difference scale by setting $\psi_1 = 0$ and $\psi_N = 1$. Then σ^2 is a separate free parameter measuring inconsistency. Alternatively, we could normalize the difference scale so that $\psi_1 = 0$ and $\psi_N = 1/\sigma$. A plot of the resulting difference scale then also includes a visual indication of the observer's inconsistency. The default method in our package uses the second method described (for discussion, see Knoblauch & Maloney 2012).

We modeled the observer's inconsistency by a Gaussian random variable; it is possible that other choices of distribution would lead to different, inconsistent estimates of the parameters $\psi_1 < \psi_2 < \cdots < \psi_N$. Maloney & Yang (2003) investigated the robustness of MLDS to failures of the distributional assumptions. If, for example, the maximum likelihood fit were based on the Gaussian assumption, but the actual distribution were very different, would the estimates of $\psi_1 < \psi_2 < \cdots < \psi_N$ be markedly in error? Surprisingly, Maloney & Yang (2003) found that, for a range of distributions, they were not. Thus, while MLDS is apparently based on a parametric (Gaussian) assumption, the assumption has little effect on the solutions obtained.

2.2. An Example

As an example, we present the experiment of Obein et al. (2004), in which they used MLDS to estimate how observers perceive gloss. Gloss is a complex physical phenomenon, as it depends on the interaction of both diffuse and specular reflections of the illumination from the surface. The pattern of light radiating from the surface to the eye depends on many factors. Given any physical measure of gloss, MLDS presents itself as a possible approach to measuring gloss appearance as physical gloss is varied. We can repeat this measurement with different viewing conditions. Given the physical changes in the stimulus with viewing conditions, does the appearance change? In particular, does the appearance of glossy surfaces change with viewing angle and whether or not the surface is viewed monocularly or binocularly?

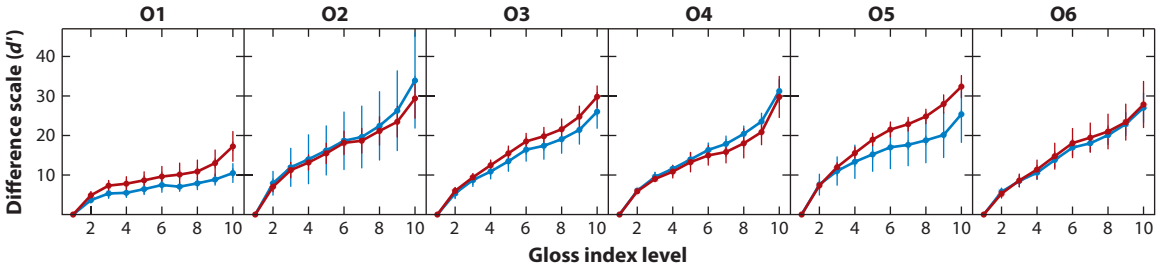

Figure 3

Average maximum likelihood difference scaling scales for six observers (O) who judged the gloss differences between pairs of pairs of stimuli differing in physical gloss for binocular (*blue*) and monocular (*red*) viewing. The error bars are ±2 standard errors of the mean. The abscissa indicates the ordinal index of the surface, not the gloss value. The ordinate values were normalized with respect to each observer's judgment noise and are thus specified as the signal detection measure d'. All of these judgments were obtained at a 60° viewing angle. Figure adapted with permission from Obein et al. (2004).

Obein et al. (2004) evaluated a set of 10 calibrated, black, coated surfaces that varied in gloss over approximately two orders of magnitude in physical gloss units, varying in appearance from matte to highly glossy. The appearance range varied with the angle of measurement (the angle between line of sight and a surface normal), and Obein et al. report measurements at 20° and 60°. They tested observers using MLDS from both of these angles of view, using both monocular and binocular vision. The study is an early application of MLDS and was unique at the time in using actual physical samples rather than stimuli generated on a computer display. As the interest in applying MLDS to more exotic and real-world varying physical continua increases, the use of MLDS with actual physical samples is increasingly common (for example, Hansmann-Roth et al. 2018).

Using real physical stimuli imposes logistical problems in running an experiment in terms of stimulus presentation, with the time needed to change the stimuli from trial to trial limiting the rate of data collection. Thus, whereas computer-driven MLDS experiments can take from 5 to 15 min for a run of 210 trials, typically depending on the number of levels and the experience of the observer, Obein et al. (2004) report that a session of 210 trials lasted 45 min.

Average MLDS scales over the four runs from the six observers tested by Obein et al. (2004) are shown in **Figure 3** for binocular and monocular viewing. As discussed above, the curves are only unique up to the addition of a constant and multiplication by a positive coefficient. In this case, we normalized each scale value by the estimated standard deviation of the judgment noise σ. We label this as d' by analogy to the measure used in signal detection theory (Green & Swets 1966) for characterizing differences in discrimination.

Observers differ in their overall judgments of magnitude of gloss differences, as shown by the different heights of the curves, although, with the exception of the first observer, the maximum heights are quite similar across observers. Given the confidence limits in the estimates, the differences between monocular and binocular viewing are surprisingly slight and unsystematic, varying between observers. The statistical reliability of monocular and binocular effects could be tested with formal model comparison, something that would be easy to do, for example, with current software (Knoblauch & Maloney 2012). What is most striking in the data is that the average shape of the curves is relatively stable between monocular and binocular viewing and also between observers. The curves display an inverted S shape that seems to be composed of two regimes. At low gloss levels, the estimates rise quickly and then begin to asymptote. The asymptotic region is followed by a second region in which the slope increases, suggesting a rapid change in perceived

glossiness. Obein et al. (2004) suggest that the low gloss levels are influenced by the lightness of the predominantly matte surface properties of the samples, while the upper segment reflects the increasingly clear specular reflectance of the illuminant. This upper segment was not previously revealed on gloss estimates using different techniques to measure appearance and using simulated samples on a computer screen.

Another interesting finding, not shown in **Figure 3**, is that, when the data from the two viewing angles, 20° and 60°, were compared with the gloss scale estimates plotted as a function of the gloss index, as in **Figure 3**, rather than the physically measured gloss units, the curves closely overlapped, suggesting that the gloss appearance was independent of the viewing angle, at least over the range explored, which would indicate some degree of gloss constancy in glossiness perception. This example shows not only that MLDS can yield reliable estimates of the appearance changes along a physical dimension, but also that the results can be used to answer pertinent questions concerning perception.

MLDS can be applied to arbitrary stimuli that can be ordered along a physical continuum. For example, Knoblauch & Maloney (2008, 2012) used the example of scatter plots that are ordered by their Pearson product moment correlation. In this case, each stimulus is a realization of a stochastic level along the continuum defined by the correlation coefficient. The reader is referred to these publications for further examples of a didactic presentation of the use of MLDS.

2.3. Review of Recent Work

Since its introduction, MLDS has been applied to measure appearance along increasingly diverse dimensions. Subsequent to its use by Obein et al. (2004) for measuring gloss, Charrier et al. (2007) compared image distortion from compression in two color spaces, showing that a color space aligned more closely with human color coding (Lab) could produce higher compression rates with less perceived distortion than one based on RGB display primaries. Rhodes et al. (2007) used MLDS to test whether observers are more sensitive to appearance changes around an average face and reported no evidence that this is the case. Knoblauch & Maloney (2008) demonstrated that perception of correlation in scatter plots follows more closely variance-accounted-for than correlation. Lindsey et al. (2010) used MLDS to estimate perceptual intervals between color stimuli to search for (and not find) evidence of categorical color boundaries. In a followup study, Brown et al. (2011) found similar evidence by relating MLDS perceptual estimates and a standard color opponent model to reaction time measurements in an effort to test the Sapir-Whorf hypothesis relating language and perception. In a series of papers, Menkovski and colleagues (Liotta et al. 2013; Menkovski & Liotta 2012; Menkovski et al. 2011a,b, 2012) tackled the application of MLDS to evaluating the quality of video image sequences, work that entailed considering whether the number and choice of samples could be selected to optimize the data collection procedure. In a series of studies, Devinck and colleagues (Devinck & Knoblauch 2012, Devinck et al. 2014, Gerardin et al. 2018a) used MLDS to quantify the long-range color filling-in of the watercolor effect (Pinna et al. 2001). An important feature of experimental design in these studies is that the stimulus feature manipulated was not the feature judged; i.e., the contour luminance was varied, but the perceived fill-in color was judged. Thus, it was necessary to introduce a control condition with the same luminance changes but with contours that did not generate a filling-in phenomenon to demonstrate that the estimated scales reflected the filling-in phenomenon under study and not an artifact of the stimulus manipulation.

Wiebel et al. (2017) compared MLDS measurements of lightness in different contexts with traditional matching measures. MLDS proved to be an effective method for comparing models of how context influences lightness perception because it is based on a statistical model that permitted

detailed simulation of different observer models. Brainard and colleagues (Brainard et al. 2018; Radonjić & Brainard 2016; Radonjić et al. 2015a,b) have adapted the MLDS paradigm to assess color shifts in color constancy experiments. Their work is focused on scaling appearance changes with respect to a focal color rather than measuring a full scale. In a more recent article, Radonjić et al. (2019) developed a variation that extends MLDS to two-dimensional physical scale values rather than a unidimensional scale (see also Haghiri et al. 2019b). MLCM, discussed below, offers a multidimensional approach to scaling when appearance depends on two or more physical scales and has much in common with MLDS.

Kingdom (2016) used MLDS to evaluate hypotheses concerning noise in contrast perception that had been suggested by previous discrimination experiments. Knoblauch et al. (2020) similarly investigated luminance and chromatic contrast response using MLDS in normal and anomalous trichromats. The estimated contrast response scales were well described by a Michaelis-Menten function, thereby allowing both response and contrast gain to be evaluated in each group. It would be of interest to extend such measurements in clinical situations to assess contrast appearance in eye disease.

By far the most adventurous applications of MLDS involve measuring the appearance of object characteristics, such as the appearance associated with simulated changes in refractive index (Fleming et al. 2011), transparency perception (Faul 2017), object solidity (Bi et al. 2018, Paulun et al. 2015), and surface texture properties (Sawayama et al. 2017). Mansour Pour et al. (2018) investigated a well-controlled class of broadband random-texture stimuli, which they called motion clouds, using MLDS and found evidence for three regimes that correspond to motion coherency, motion transparency, and motion incoherency.

A still-underdeveloped area wherein MLDS shows promise is relating these perceptual scales to functional cerebral imaging data. As the MLDS decision process is based on a signal detection model, the estimated scales can be expressed in terms of a signal-to-noise ratio, making them simple to relate to noisy neurally correlated signals. In an early study, Yang et al. (2007) evaluated cortical sites involved in processing the same physical scale when observers performed different perceptual judgments, for example, stimulus area versus stimulus color. Bellot et al. (2016) compared contrast response in subcortical and cortical structures to MLDS-derived contrast scales in different age groups. Age influenced both response and contrast gain of the MLDS-derived response functions. Age effects on contrast response were also detected in the lateral geniculate nucleus, superior colliculus, and cortical area V1. Finally, Gerardin et al. (2018a) used MLDS-estimated appearance of watercolor filling-in to help identify the cortical areas that best decoded the perceptual filling-in versus the presentation of a matching uniform chromaticity. Another interesting direction is provided by Haghiri et al. (2019a), who used crowdsourcing via Amazon Mechanical Turk to obtain a much larger number of observers performing MLDS in an efficient manner than would typically be possible in a laboratory environment.

3. MAXIMUM LIKELIHOOD CONJOINT MEASUREMENT

3.1. The Model

Additive conjoint measurement permits one simultaneously to measure and model the contributions of multiple physical dimensions to measured appearance. The model in its simplest form is additive, and we confine attention to additive conjoint measurement. Each stimulus is characterized by physical measures on each of several dimensions. The observer's task is to rank the stimuli by some dependent, perceptual measure. The experimenter wishes to determine how the physical dimensions contribute to judged perceptual appearance.

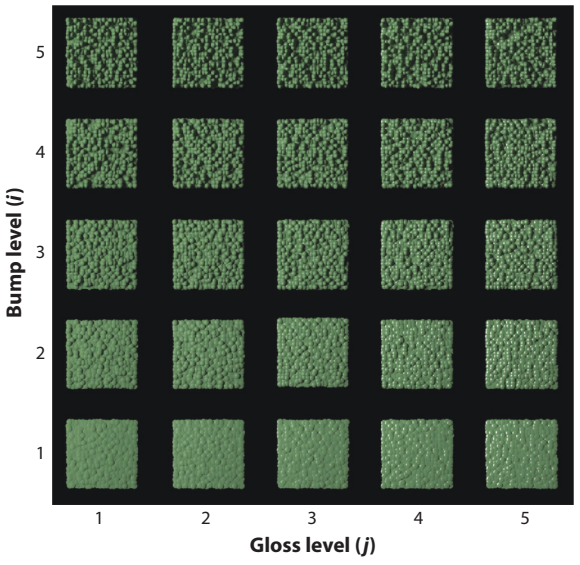

Figure 4

A five-by-five array of rendered surfaces differing in physical glossiness and a measure of physical roughness. Each column has constant glossiness and five different levels of roughness that decrease from the top to the bottom of the column. Each row has constant roughness and five different levels of glossiness that increase from left to right. Figure adapted with permission from Ho et al. (2008).

Figure 4, for example, contains a five-by-five array of rendered surfaces (stimuli) taken from Ho et al. (2008). The 25 stimuli vary in two physical dimensions, roughness and glossiness. (Ho et al. use the term bumpiness rather than roughness, but the latter is more common in the literature, and we adopt it in this review).

In one experiment, observers judged the roughness of surfaces; in a second experiment, a different set of observers viewed the same stimuli and judged which of each pair was glossier. We emphasize that the stimuli in the two experiments were identical but that observers were instructed to extract a perceptual analog of each physical dimension in turn while squelching any effect of the other.

As an application of additive conjoint measurement, the experiment of Ho et al. (2008) is unusual in two respects. First, they assessed how well observers could judge a single physical dimension (gloss or roughness) when both stimulus dimensions were varied jointly, a perceptual constancy (more precisely, two perceptual constancies). They sought to model the interaction (interference) between dimensions: to what degree did changes in physical roughness affect perceived gloss and vice versa. Second, the judgments of appearance in the roughness experiment and in the glossiness experiment were the perceptual correlates of one of the physical dimensions varied to produce stimuli. This need not be the case in other applications of MLCM.

We do not present the study in detail, but instead use it to outline the issues that need to be considered in using conjoint measurement. The first is that the physical scales are simply arbitrary. The glossiness parameter was taken from a standard rendering package. The roughness parameter was the variance of surface distance variation along the line of sight. A possible approach to modeling such data would be applying MLDS twice to the isolated scales before attempting to test (in some manner) how they interact.

MLCM will determine separate scales for the physical dimensions as part of the process of fitting its model. We denote the first physical scale as φ^1, the second as φ^2, etc. The sampled values on each scale are denoted by subscripts. For the dth dimension, these would be $\varphi^d_1 < \cdots < \varphi^d_{N_d}$, where the inequality refers to the ordering on the physical scale, and N_d is the number of sampled values on the dth physical scale.

Second, in MLCM, the focus is not solely on the physical scales, but also on characterizing the interaction between scales. To simplify notation, let us assume that there are only two physical dimensions, $d = 2$.

Any stimulus in **Figure 4** is specified by its roughness and glossiness levels, and we specify stimuli by ordered pairs: $[\varphi^1_j, \varphi^2_k]$ denotes the stimulus that had level j of dimension 1 and level k on dimension 2. There are N_1N_2 stimuli (25 in **Figure 4**). We can simplify notation by replacing each stimulus $[\varphi^1_j, \varphi^2_k]$ by an ordered pair $[j, k]$. The psychophysical task is to order any two stimuli (ordered pairs) in whatever measure of appearance the experimenter chooses.

The model is similar to that of MLDS but expanded to two or more dimensions. To each physical measure φ^1_j we associate a measure ψ^1_j that is interpreted as the contribution to the appearance measure of that level of the first dimension. The physical measure φ^2_k is mapped to ψ^2_k, and if there are only two physical dimensions, then we define the appearance magnitude as $\psi^1_j + \psi^2_k$, the sum of the contributions from each dimension. The model is readily extended to more than two dimensions. As in MLDS, we formulate a model of the observer's judgments based on these appearance magnitudes for a nonstochastic observer and then for a fallible, stochastic observer. First, given any two stimuli $[i, j]$ and $[k, l]$, we require that

$$\psi^1_i + \psi^2_j < \psi^1_k + \psi^2_l \text{ if and only if } [i, j] \prec [k, l], \qquad 3.$$

a simple additive model.

The number of possible stimuli is N_1N_2 in the two-dimensional case. In the work of Ho et al. (2008), $N_1 = N_2 = 5$, and the number of possible stimuli is 25, as shown in **Figure 4**. At first glance, there are $N_1 + N_2$ free parameters for the MLCM model: $\psi^1_1, \psi^1_2, \ldots, \psi^1_{N_1}, \psi^2_1, \psi^2_2, \ldots, \psi^2_{N_2}$. However, as with MLDS, adding a constant to all the parameters or multiplying all the parameters by a positive constant leads to no change in the predictions of the model. There are effectively $N_1 + N_2 - 2$ free parameters, and we can arbitrarily normalize the scales. For example, we could set $\psi^2_{N_2} = 1$ for convenience.

The model presented above assumes that the observer's responses are not stochastic: Presented with the same stimuli, the observer makes the same response. We remedy this defect by including a stochastic component in the observer's judgment:

$$\psi^1_i + \psi^2_j < \psi^1_k + \psi^2_l + \epsilon \text{ if and only if } [i, j] \prec [k, l], \qquad 4.$$

where ϵ is a Gaussian random variable with mean 0 and variance σ^2. The additional parameter σ^2 brings the total of free parameters back to $N_1 + N_2 - 1$ (for details of the maximum likelihood fitting procedure, see Knoblauch & Maloney 2012).

An experimental design could require the observers to compare all stimuli $[ij]$ to all distinct stimuli $[kl]$. For the 25 stimuli in **Figure 4**, the observer could make 300 distinct forced-choice comparisons. If the full set of comparisons is repeated one or more times, then the number of trials increases accordingly. If the number of levels is larger, then fits based on a subset of the samples would be acceptable (Abbatecola et al. 2020). If stimuli are chosen at random, then the critical variable seems to be the total number of trials tested (Abbatecola et al. 2020, Knoblauch & Maloney 2012).

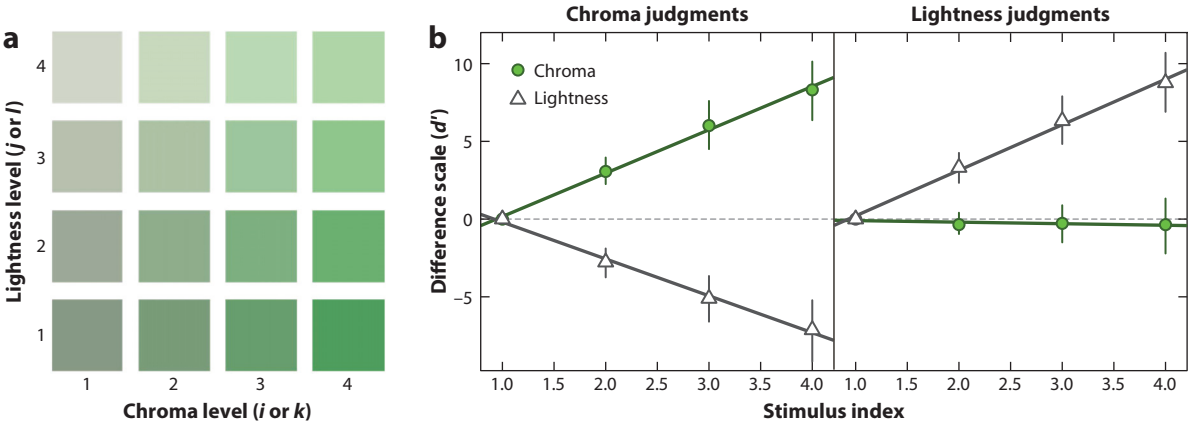

Figure 5

(*a*) Representation of set of stimuli varying in chroma and lightness for one of four hues tested by Rogers et al. (2016) in a maximum likelihood conjoint measurement experiment in which observers compared either the chroma or the lightness of a stimulus pair randomly chosen from the set. On each trial, a stimulus of chroma *i* and lightness *j* was compared with another of chroma *k* and lightness *l*. The axis units indicate indices assigned to each stimulus. (*b*) Estimated contributions of chroma (*green circles*) and lightness (*white triangles*) to chroma judgments (*left*) and lightness judgments (*right*) under an additive model fit, replotted from Rogers et al. (2016). Points are averages of the results from 15 observers each fit by maximum likelihood. The error bars are 95% confidence intervals. The lines are the fixed-effect or population estimates from a generalized linear mixed-effects model fit to the responses of all observers with the random effect attributed to observer variation of the slope for each component. Figure adapted with permission from Rogers et al. (2016).

3.2. An Example

As described in the previous section, the seminal article introducing MLCM was that of Ho et al. (2008), in which the mutual influences of surface gloss and roughness were investigated in rendered stereoscopically fused images. An extended tutorial using this work as an example was detailed by Knoblauch & Maloney (2012), so we develop a different example in this review, the experiment of Rogers et al. (2016) in which they used MLCM to measure how lightness influences chroma judgments and chroma influences lightness judgments. They tested four hues (red, yellow, green, and blue) with stimuli selected from a matrix of four levels each of lightness and chroma, yielding 16 stimuli and 128 unordered pairs, including self-comparisons between stimuli whose levels were the same on each dimension. These self-comparisons were not included in the analyses, as they do not contribute to the estimate of the scale values, but they can be evaluated separately to test for possible response biases.

The stimuli were specified in Commission Internationale de l'Eclairage (CIE) LCh_{uv} space, a cylindrical version of CIELUV, where L is the lightness, C_{uv} is the chroma, and h_{uv} is the hue (Poynton 2012). An example of the stimulus matrix for the green hue is shown in **Figure 5*a***. Thirty observers were tested in the experiment, with 15 participating in experiments judging red and yellow stimuli and another 15 in experiments judging blue and green. Each chroma/lightness matrix was tested in eight separate sessions, with sessions divided between conditions in which observers judged either the chroma or lightness on a given trial. Specifically, in a lightness session, observers chose which of the presented pair appeared lighter, and in a chroma session, observers judged which stimulus had a stronger chroma. For example, for the green hue judgments, the observers chose the stimulus patch that appeared greener. It is important to note that, for a given hue, the stimulus set presented to the observer was exactly the same in the two conditions, and only the instructions as to which stimulus feature to judge differed.

The average results of the 15 observers for the green hue conditions fit under the additive model are shown in **Figure 5*b***. When the observers judged chroma, the chroma component increased with the chroma of the stimulus, but the lightness component decreased with the lightness of the stimulus. This means that increasing the stimulus lightness was more likely to lead the observer to report a stimulus as less green. The decrease in chroma with increase of lightness is referred to as veiling (Krantz 1975). The same phenomenon was observed for blue and red hues, but for yellow, the chroma component was nearly independent of the lightness component. Increasing lightness slightly increased the yellow appearance of the stimulus.

Figure 5*b* shows the chroma and lightness contributions to lightness judgments. In this case, increasing stimulus lightness led to increasing lightness judgments but independent of the chroma level of the stimulus. Again, similar results appeared for the other hues, except for red, for which there was a small but significant positive contribution of chroma to lightness.

An interesting feature of this study is that, in a pilot experiment, each of the hues at fixed lightness and an achromatic series with chroma set to 0 were evaluated with MLDS, each being tested at 10 levels. The four levels presented in the MLCM experiment were selected to have equal perceptual differences based on the MLDS results. This choice of stimulus levels tested is what led to the linear variation of each component with its levels in **Figure 5*b***. The linearity resulting from this choice of stimulus levels, in addition, permitted the data to be simply fit with a generalized linear mixed-effects model (GLMM) in which, given the identifiability constraint that the curves pass through 0 at the lowest values, the only parameters necessary to estimate were the slopes of the two components, thereby reducing the complexity of the model (Bates et al. 2015, Knoblauch & Maloney 2012). The solid lines in **Figure 5*b*** correspond to the population estimates from the additive GLMM model (for details, see Rogers et al. 2016). The results support the claim that chroma and lightness contribute in an additive fashion to chroma judgments but that lightness judgments do not depend on the chroma of the stimulus patch. Lisi & Gorea (2016), Chammat et al. (2011), and Nichiporuk et al. (2017, 2018) further exploited the linearity of the variation of components in an MLCM experiment to fit their data using a GLMM.

3.3. Review of Recent Work

Since Ho et al. (2008) introduced MLCM as a method for investigating surface-related properties, several studies have used it to explore object-related perceptual dimensions. Hansmann-Roth et al. (2018) manufactured real stimuli to test the influence of glossiness and shape on one another. Hansmann-Roth & Mamassian (2017) also investigated contextual influences on gloss perception using MLCM. For MLCM, an initial focus is typically directed to testing three nested models, independence, additivity, and a full model, the latter evaluating the possibility of interaction effects between the dimensions tested (also called the saturated model because the maximum number of identifiable parameters is estimated). Emrith et al. (2010) explored the role of higher-order statistics in surface texture perception. Interaction effects were significant, and Emrith et al. demonstrated how to visualize these effects.

MLCM was further used to explore the perception of scattering and diffusion in translucence perception. Chadwick et al. (2018) used mixtures of concentrations of tea and milk as the physical dimensions with real and simulated images. Differences between the results for the two sets of images point to cues in the real images that the simulations did not capture. The interesting approach of these authors was to explore whether decision models based on particular image heuristics were sufficient to account for the observers' behavior in these experiments, leading to propositions on the computations that the visual system might employ in translucence perception.

In their studies of the watercolor effect, Gerardin et al. (2014) used MLCM to investigate all the pairings of three different dimensions that might influence the phenomenon. As in their previous studies of this phenomenon using MLDS, judgments of control stimuli allowed them to conclude that observers' judgments were based on the hue appearance of filling-in, and not the variations in stimulus dimensions. The additivity model best described the data. As expected, therefore, the contribution of a given dimension did not depend on the other dimension with which it was paired. In a further MLCM study, the influence of background luminance was paired with contour luminance, and observers judged either the hue or the brightness of the filling-in (Gerardin et al. 2018b). In this case, interaction effects were significant, but the two judgments depended on background luminance in different fashions. Increasing background luminance generated assimilation based on the hue judgments but contrast based on the brightness judgments.

MLCM is based on paired comparisons. This creates possibilities for applications in which the choice is based on preference responses, which might more easily be recorded from nonverbal organisms. As an example, Rogers et al. (2018) followed up their initial work on chroma and lightness perception in adults (Rogers et al. 2016; described in the previous section) with a study in infants in which a forced-choice preferential-looking paradigm (Teller 1979) was used. Infants' first looking response toward one of the stimuli comprising the pair was recorded as the choice based on analyses of videos obtained during the session. The procedure was simplified for the limited attentional resources of infants by reducing the stimuli to a 3×3 set, yielding only 36 unique unordered pairs. Each pair was intended to be presented twice, with the left to right order reversed on the second presentation, but only one-quarter of the 21 infants tested completed all 72 trials. Nevertheless, using a GLMM, the data from all infants could be analyzed to test the three nested hypotheses and to demonstrate the contributions of chroma and lightness to the infants' choices. Unlike the adult experiments, the infants cannot be instructed to compare the chroma or the lightness in a stimulus pair; the judgments are best interpreted as resulting from the salience of the components in the stimuli. In control experiments, adults who were asked to make a saccade to the most salient stimulus performed quite differently than did the infants, and the results were more variable across observers, suggesting a range of strategies in the adults. Nevertheless, the infant preferences generated contributions of chroma and lightness to salience that qualitatively were most similar to the case in the previous paper when adults were asked to make chroma judgments.

In an original application, Lisi & Gorea (2016) used MLCM to demonstrate a time constancy phenomenon analogous to size constancy. When, and only when, observers were provided information about viewing distance, jointly modulated size and speed had no influence on duration perception. Chammat et al. (2011) used MLCM to demonstrate that emotional intensity of natural images influences perceived contrast. MLCM has also been used to confirm a race bias on lightness perception of faces (Nichiporuk et al. 2017, 2018). An interesting technical aspect of these studies is the demonstration that the stimulus matrix need not be square. Nichiporuk et al.'s (2017, 2018) stimulus design included two levels of race (Caucasian and African) and 13 levels of face lightness.

4. FUTURE DIRECTIONS

4.1. Maximizing Likelihood

MLDS and MLCM link human perceptual judgments to parametric stochastic models of the judgment process. In both methods, each trial is modeled as an independent Bernoulli random variable, and the probability of the outcome of the trial is determined by the settings of the model

parameters. [Knoblauch & Maloney (2012) describe the fitting procedure and the underlying statistical theory in detail.] In our approach, the key step in fitting is to select the settings of the model parameters that maximize the overall probability (likelihood) of the observed responses, maximum likelihood estimation (MLE). MLE has many useful properties. MLE estimates are asymptotically unbiased and uniformly minimum variance and can be used with standard model comparison methods (Knoblauch & Maloney 2012).

The fitting procedures for MLDS and MLCM are equivalent to fitting generalized linear models (GLMs), and we can use GLM fitting methods to compute MLDS and MLCM solutions. These methods are also based on maximizing likelihood, and using GLM methods is a convenience. The underlying MLDS or MLCM model is not changed.

There are alternative approaches to maximizing likelihood. Schneider and colleagues (Schneider 1980a,b; Schneider et al. 1974), for example, used a model similar to what we refer to above as the nonstochastic model and chose model parameter settings that minimized the count of discrepancies between model predictions and the observer's actual responses. As σ^2 (the variance of the judgment uncertainty, ϵ) approaches 0, their fitting method converges to MLDS. In most applications, though, σ^2 is substantially greater than 0.

4.2. Numerical Optimization

The key step in MLDS is the maximization of likelihood using numerical maximization methods. There exist many such methods, and one could create a variant of MLDS simply by changing from one optimization method to another. Knoblauch & Maloney (2008, 2012) provide two standard maximization methods (one based on GLM) in their MLDS package. While some optimization methods may prove superior to others for particular kinds of problems, such changes do not alter the basic theory underlying MLDS. At best, they decrease the probability of spurious solutions (local minima), to which all optimization methods are prone. Still, the basic question remains: Which optimization methods provide the best solutions for MLDS and MLCM in specific applications? Some research has been done on this issue recently (Haghiri et al. 2019b).

Directions for future research would include development of more powerful and reliable optimization methods for MLDS and MLCM and investigation of the limits of existing measures.

4.3. Alternatives to Maximum Likelihood Difference Scaling and Maximum Likelihood Conjoint Measurement

In this section, we describe some alternatives to MLDS and MLCM.

4.3.1. Direct scaling. A common alternative to MLDS is to use numerical estimates of appearance, an approach referred to as direct scaling and favored by Stevens (1946). The observer is shown stimuli $\varphi_1, \ldots, \varphi_N$ one at a time and asked to assign a numerical estimate of magnitude to each. There is no reason to anticipate that the resulting scale values can be used to predict differences or anything else other than the estimates themselves. It is based on the assumption that human use of numerical ratings is readily interpretable. In an elegant article, Augustin (2006) discusses possible interpretations of ratings obtained by direct scaling and what each interpretation presupposes. Schneider and colleagues (Schneider 1980a,b; Schneider et al. 1974) compared direct scaling to a difference scaling method that is similar to MLDS but with a different statistical framework. They found that their difference scaling estimates were more stable and readily interpretable. [The reader is also referred to Wiebel et al. (2017), who compared matching and MLDS in lightness judgments.]

4.3.2. Concatenated just noticeable differences. Another alternative is to develop scales based on concatenation of just noticeable differences (JNDs). The experimenter uses threshold measurements to create a list of physical stimuli $\varphi_1 < \cdots < \varphi_N$, with adjacent stimuli being equidiscriminable. That is, the probability of judging which of two adjacent stimuli $\varphi_i < \varphi_{i+1}$ is more intense is a fixed value π. The difference is referred to as a JND or, more precisely, a π-JND. The resulting scale resembles a Thurstone Scale, Case V (Thurstone 1927), and this method is sometimes referred to as Thurstone scaling.

A celebrated claim is that stimuli that are a certain number of JNDs apart differ equally in appearance. This claim originated with Fechner (1860) and is the basis for Fechner's Law. The central question is whether concatenated JNDs do, in fact, capture differences in appearance. This claim is controversial (Stevens 1961), with mixed results in the literature (Laming & Laming 1996; Schneider 1980a,b; Schneider et al. 1974). The method of concatenated JNDs is potentially time consuming; the number of forced-choice trials needed to measure concatenated JNDs can be very large compared to the number of forced-choice trials needed for MLDS. Moreover, it is clear what the difference scale offers: prediction of perceived differences between stimuli on a continuum. The interpretation of concatenated JNDs is less obvious.

A scale based on concatenated JNDs is based, in effect, on very local, near-threshold information, the rate of confusion between nearby stimuli. If two stimuli on the scale are suprathreshold in the sense that the rate of confusion is too small to estimate reliably, then judgments of these stimuli will not appreciably constrain the concatenated JND scale. MLDS, in contrast, includes comparison of large scale differences, the intervals whose differences are compared. One might say that MLDS depends on more global properties of appearance.

Similar issues arise with MLCM: The key step in MLCM is to develop scales for each of the physical dimensions and use sums of the ratings on these scales to predict observers' judgments in ordering stimuli. We emphasize that MLCM is not simply application of MLDS to multiple dimensions. It includes an assessment of how the dimensions interact and, in effect, places the dimensions on a common scale. Indeed, scaling the MLCM dimensions separately by repeated applications of MLDS allows us to test the hypothesis that the scales arising from the two methods are the same. We view these and related questions as fundamentally empirical. MLDS and MLCM provide well-defined methods with clear interpretations for measuring specified aspects of experience based on specific linking hypotheses with experience. It is the developer's burden to demonstrate that a novel method measures anything at all and to demonstrate that it measures what it is claimed to measure. Experimentally, we can and should test and verify these claims—or reject them.

Similarly, there are other methods that provide partial information about appearance, some specific to a single domain. In color, for example, we have color matching (Kaiser & Boynton 1996) and opponent hue cancellation (Hurvich 1981). Both methods tell us something about appearance, but neither allows us to estimate a scale of appearance comparable to those provided by MLDS and MLCM, and we consider them no further.

4.4. Diagnostics

For the stimuli in **Figure 1**, it is plausible that observers can order intervals. Given other possible sets of physical stimuli (e.g., roughness), it is not evident that an observer's judgments will make any sense at all. There are many ways for a surface to be rough. The observer may simply be guessing or basing his judgments on a model other than the one that the experimenter is considering. If an observer were to judge that $ij \prec kl$ and $kl \prec mn$ but $ij \succ mn$, then we would hesitate to ascribe any sort of scale to the set of judgments.

In addition to specifying a psychophysical judgment and a model of that judgment, we need to specify tests of the assumptions underlying the model and the linking hypothesis itself. We need to be able to assess failures of models when the model is not appropriate for the data. In fitting statistical models to data, it is common practice to test the assumptions of each fitted model, not just the overall goodness of fit. We look for patterned failures of the model that call into question its appropriateness. In statistics, these tests are referred to as diagnostics, and diagnostics such as the runs test (Wood 2017) are commonly employed in multiple regression (Belsley et al. 1980). Knoblauch & Maloney (2008) discuss possible diagnostic statistics for MLDS; further work is needed.

4.5. New Models and Methods

The key step in MLDS and MLCM is to link a perceptual judgment with a model of the judgment. There is a wealth of possible models to choose among, including those described in the multivolume *Foundations of Measurement* (Krantz et al. 1971, 1989; Suppes et al. 1990). Whether such models are useful is an empirical question that remains to be tested. We briefly mention polynomial conjoint measurement, an extension of additive conjoint measurement that allows for both multiplicative and additive interactions among physical dimensions. Logvinenko & Maloney (2006) applied the model without labeling it as such to analyze rated similarity between surfaces in scenes differing in illumination. The judgment that they employed was a form of asymmetric lightness matching, with observers asked to rate the dissimilarity of surface lightness of two surfaces embedded in simple scenes differing in illumination.

If MLDS or MLCM captures human data, then it is very clear what has been achieved. The MLDS observer can order perceptual intervals in a way that is captured by a parsimonious model. A similar claim can be made for MLCM. The assumptions underlying both methods are explicit (Knoblauch & Maloney 2012) and testable. If MLDS or MLCM does not capture human data, then we hope that diagnostic tests detect the inconsistencies between human data and the models.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank David H. Brainard for critical comments. L.T.M. was supported by the National Institutes of Health (National Eye Institute, grant EY019889), the Guggenheim Foundation, and the Institute for Advanced Studies of Paris. K.K. was supported by the Agence Nationale de la Recherche (grants ANR-15-CE32-0016 CORNET, ANR-17-NEUC-0004 A2P2MC, ANR-17-HBPR-0003 CORTICITY, and ANR-19-CE37-025 DUAL_STREAMS).

LITERATURE CITED

Abbatecola C, Beneyton K, Gerardin P, Kennedy H, Knoblauch K. 2020. Voice and face gender perception engages multimodal integration via multiple feedback pathways. bioRxiv 884668. **https://doi.org/10.1101/2020.01.07.884668**

Anderson D, Burnham K. 2004. *Model Selection and Multi-Model Inference*. Berlin: Springer. 2nd ed.

Augustin T. 2006. Stevens' direct scaling methods and the uniqueness problem. *Psychometrika* 71:469–81

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1):1–48

Bellot E, Coizet V, Warnking J, Knoblauch K, Moro E, Dojat M. 2016. Effects of aging on low luminance contrast processing in humans. *NeuroImage* 139:415–26

Belsley DA, Kuh E, Welsch R. 1980. *Identifying Influential Data and Sources of Collinearity*. New York: Wiley

Bi W, Jin P, Nienborg H, Xiao B. 2018. Estimating mechanical properties of cloth from videos using dense motion trajectories: human psychophysics and machine learning. *J. Vis.* 18:12

Brainard DH, Cottaris NP, Radonjić A. 2018. The perception of colour and material in naturalistic tasks. *Interf. Focus* 8:20180012

Brindley GS. 1970. *Physiology of the Retina and Visual Pathway*. London: Edward Arnold. 2nd ed.

Brown AM, Lindsey DT, Guckes KM. 2011. Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *J. Vis.* 11:2

Chadwick AC, Cox G, Smithson HE, Kentridge RW. 2018. Beyond scattering and absorption: perceptual unmixing of translucent liquids. *J. Vis.* 18:18

Chammat M, Jouvent R, Dumas G, Knoblauch K, Dubal S. 2011. Interactions between luminance contrast and emotionality in visual pleasure and contrast appearance. *Percept. ECVP Abstr.* 40:22

Charrier C, Maloney LT, Cherifi H, Knoblauch K. 2007. Maximum likelihood difference scaling of image quality in compression-degraded images. *J. Opt. Soc. Am. A* 24:3418–26

Devinck F, Gerardin P, Dojat M, Knoblauch K. 2014. Spatial selectivity of the watercolor effect. *J. Opt. Soc. Am. A* 31:1–6

Devinck F, Knoblauch K. 2012. A common signal detection model accounts for both perception and discrimination of the watercolor effect. *J. Vis.* 12:19

Efron B, Tibshirani RJ. 1994. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press

Emrith K, Chantler MJ, Green PR, Maloney LT, Clarke AD. 2010. Measuring perceived differences in surface texture due to changes in higher order statistics. *J. Opt. Soc. Am. A* 27:1232–44

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–33

Faul F. 2017. Toward a perceptually uniform parameter space for filter transparency. *ACM Trans. Appl. Percept.* 14:13

Fechner G. 1860. *Elemente der Psychophysik*. Leipzig, Ger.: Breitkopf & Härtel

Fleming RW, Jakel F, Maloney LT. 2011. Visual perception of thick transparent materials. *Psychol. Sci.* 22:812–20

Fleming RW, Nishida S, Gegenfurtner KR. 2015. Perception of material properties. *Vis. Res.* 115:157–62

Geisler WS. 1989. Sequential ideal-observer analysis of visual discriminations. *Psychol. Rev.* 96:267–314

Gerardin P, Abbatecola C, Devinck F, Kennedy H, Dojat M, Knoblauch K. 2018a. Neural circuits for long-range color filling-in. *NeuroImage* 181:30–43

Gerardin P, Devinck F, Dojat M, Knoblauch K. 2014. Contributions of contour frequency, amplitude, and luminance to the watercolor effect estimated by conjoint measurement. *J. Vis.* 14:9

Gerardin P, Dojat M, Knoblauch K, Devinck F. 2018b. Effects of background and contour luminance on the hue and brightness of the watercolor effect. *Vis. Res.* 144:9–19

Green DM, Swets JA. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley. 1st ed.

Haghiri S, Wichmann F, von Luxburg U. 2019a. Comparison-based framework for psychophysics: lab versus crowdsourcing. arXiv:1905.07234 [cs.LG]

Haghiri S, Wichmann F, von Luxburg U. 2019b. Estimation of perceptual scales using ordinal embedding. arXiv:1908.07962 [cs.LG]

Hansmann-Roth S, Mamassian P. 2017. A glossy simultaneous contrast: conjoint measurements of gloss and lightness. *i-Perception* 8. **https://doi.org/10.1177/2041669516687770**

Hansmann-Roth S, Pont SC, Mamassian P. 2018. Contextual effects in human gloss perception. *Electron. Imag.* 2018:1–7

Hillis JM, Brainard DH. 2005. Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *J. Opt. Soc. Am. A* 22:2090–106

Hillis JM, Brainard DH. 2007a. Distinct mechanisms mediate visual detection and identification. *Curr. Biol.* 17:1714–19

Hillis JM, Brainard DH. 2007b. Do common mechanisms of adaptation mediate color discrimination and appearance? Contrast adaptation. *J. Opt. Soc. Am. A* 24:2122–33

Ho YX, Landy MS, Maloney LT. 2008. Conjoint measurement of gloss and surface texture. *Psychol. Sci.* 19:196–204
Hurvich LM. 1981. *Color Vision*. Sunderland, MA: Sinauer Assoc.
Kaiser PK, Boynton RM. 1996. *Human Color Vision*. Washington, DC: Opt. Soc. Am. 2nd ed.
Kingdom FA. 2016. Fixed versus variable internal noise in contrast transduction: the significance of Whittle's data. *Vis. Res.* 128:1–5
Kingdom FA, Prins N. 2016. *Psychophysics: A Practical Introduction*. London: Academic. 2nd ed.
Knoblauch K, Maloney L. 2008. MLDS: maximum likelihood difference scaling in R. *J. Stat. Softw.* 25(2):1–26
Knoblauch K, Maloney LT. 2012. *Modeling Psychophysical Data in R*. Berlin: Springer
Knoblauch K, Maloney LT, Aguilar G. 2019. MLCM: maximum likelihood conjoint measurement. *R Package Version 0.4.2*. **https://CRAN.R-project.org/package=MLCM**
Knoblauch K, Marsh-Armstrong B, Werner JS. 2020. Suprathreshold contrast response in normal and anomalous trichromats. *J. Opt. Soc. Am. A* 37:133–44
Krantz DH. 1975. Color measurement and color theory: II. Opponent-colors theory. *J. Math. Psychol.* 12:304–27
Krantz DH, Luce RD, Suppes P, Tversky A. 1971. *Foundations of Measurement*, Vol. 1: *Additive and Polynomial Representations*. London: Academic
Krantz DH, Luce RD, Suppes P, Tversky A. 1989. *Foundations of Measurement*, Vol. 2: *Geometric, Threshold, and Probabilistic Representations*. London: Academic
Laming J, Laming D. 1996. J. Plateau: on the measurement of physical sensations and on the law which links the intensity of these sensations to the intensity of the source. *Psychol. Res.* 59:134–44
Landy MS, Maloney LT, Johnston EB, Young M. 1995. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vis. Res.* 35:389–412
Lindsey DT, Brown AM, Reijnen E, Rich AN, Kuzmova YI, Wolfe JM. 2010. Color channels, not color appearance or color categories, guide visual search for desaturated color targets. *Psychol. Sci.* 21:1208–14
Liotta A, Mocanu DC, Menkovski V, Cagnetta L, Exarchakos G. 2013. Instantaneous video quality assessment for lightweight devices. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, pp. 525–31. New York: ACM
Lisi M, Gorea A. 2016. Time constancy in human perception. *J. Vis.* 16:3
Logvinenko AD, Maloney LT. 2006. The proximity structure of achromatic surface colors and the impossibility of asymmetric lightness matching. *Percept. Psychophys.* 68:76–83
Maloney LT, Yang JN. 2003. Maximum likelihood difference scaling. *J. Vis.* 3:573–85
Maloney LT, Zhang H. 2010. Decision-theoretic models of visual perception and action. *Vis. Res.* 50:2362–74
Mansour Pour K, Gekas N, Perrinet L, Mamassian P, Montagnini A, Masson G. 2018. Speed uncertainty and motion perception with naturalistic random textures. *J. Vis.* 18:345
McCourt ME, Blakeslee B. 1994. Contrast-matching analysis of grating induction and suprathreshold contrast perception. *J. Opt. Soc. Am. A* 11:14–24
Menkovski V, Exarchakos G, Liotta A. 2011a. Adaptive testing for video quality assessment. *Proceedings of the 2nd International Workshop on Future Television (EuroITV 2011), June 29*, ed. MJ Dama'sio, G Cardoso, C Quico, D Geerts, pp. 128–31. Lisbon: Univ. Lusófona Humanid. Tecnol.
Menkovski V, Exarchakos G, Liotta A. 2011b. The value of relative quality in video delivery. *J. Mobile Multimed.* 7:151–62
Menkovski V, Exarchakos G, Liotta A. 2012. Tackling the sheer scale of subjective QoE. In *Mobile Multimedia Communications*, ed. L Atzori, J Delgado, D Giusto, pp. 1–15. Berlin: Springer
Menkovski V, Liotta A. 2012. Adaptive psychometric scaling for video quality assessment. *Image Commun.* 27:788–99
Nichiporuk N, Knoblauch K, Abbatecola C, Shevell S. 2017. The lightness distortion effect: Additive conjoint measurement shows race has a larger influence on perceived lightness of upright than inverted faces. *J. Vis.* 17:245
Nichiporuk N, Knoblauch K, Abbatecola C, Shevell S. 2018. Does observer's ethnicity affect perceived face lightness? A study of the face-lightness distortion effect for African American and Caucasian observers. *J. Vis.* 18:1099

Nishida S, Kawabe T, Sawayama M, Fukiage T. 2018. Motion perception: from detection to interpretation. *Annu. Rev. Vis. Sci.* 4:501–23
Obein G, Knoblauch K, Vienot F. 2004. Difference scaling of gloss: nonlinearity, binocularity, and constancy. *J. Vis.* 4:711–20
Paulun VC, Kawabe T, Nishida S, Fleming RW. 2015. Seeing liquids from static snapshots. *Vis. Res.* 115:163–74
Pelli DG, Bex P. 2013. Measuring contrast sensitivity. *Vis. Res.* 90:10–14
Pinna B, Brelstaff G, Spillmann L. 2001. Surface color from boundaries: a new "watercolor" illusion. *Vis. Res.* 41:2669–76
Poynton C. 2012. *Digital Video and HD: Algorithms and Interfaces*. Amsterdam: Elsevier
R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna: R Found. Stat. Comput.
Radonjić A, Brainard DH. 2016. The nature of instructional effects in color constancy. *J. Exp. Psychol. Hum. Percept. Perform.* 42:847–65
Radonjić A, Cottaris NP, Brainard DH. 2015a. Color constancy in a naturalistic, goal-directed task. *J. Vis.* 15:3
Radonjić A, Cottaris NP, Brainard DH. 2015b. Color constancy supports cross-illumination color selection. *J. Vis.* 15:13
Radonjić A, Cottaris NP, Brainard DH. 2019. The relative contribution of color and material in object selection. *PLOS Comput. Biol.* 15:e1006950
Rhodes G, Maloney LT, Turner J, Ewing L. 2007. Adaptive face coding and discrimination around the average face. *Vis. Res.* 47:974–89
Roberts FS. 1985. *Measurement Theory*. Cambridge, UK: Cambridge Univ. Press
Rogers M, Franklin A, Knoblauch K. 2018. A novel method to investigate how dimensions interact to inform perceptual salience in infancy. *Infancy* 23:833–56
Rogers M, Knoblauch K, Franklin A. 2016. Maximum likelihood conjoint measurement of lightness and chroma. *J. Opt. Soc. Am. A* 33:A184–93
Sawayama M, Nishida S, Shinya M. 2017. Human perception of subresolution fineness of dense textures based on image intensity statistics. *J. Vis.* 17:8
Schneider B. 1980a. Individual loudness functions determined from direct comparisons of loudness intervals. *Percept. Psychophys.* 28:493–503
Schneider B. 1980b. A technique for the nonmetric analysis of paired comparisons of psychological intervals. *Psychometrika* 45:357–72
Schneider B, Parker S, Stein D. 1974. The measurement of loudness using direct comparisons of sensory intervals. *J. Math. Psychol.* 11:259–73
Stevens S. 1946. On the theory of scales of measurement. *Science* 103:677–80
Stevens SS. 1961. To honor Fechner and repeal his law. *Science* 133:80–86
Suppes P, Krantz DH, Luce RD, Tversky A. 1990. *Foundations of Measurement*, Vol. 3: *Representation, Axiomatization, and Invariance*. London: Academic
Takasaki H. 1966. Lightness change of grays induced by change in reflectance of gray background. *J. Opt. Soc. Am.* 56:504–9
Teller DY. 1979. The forced-choice preferential looking procedure: a psychophysical technique for use with human infants. *Infant Behav. Dev.* 2:135–53
Thurstone LL. 1927. A law of comparative judgment. *Psychol. Rev.* 34:273–86
Ward F, Boynton RM. 1974. Scaling of large chromatic differences. *Vis. Res.* 14:943–49
Whittle P. 1992. Brightness, discriminability and the "crispening effect." *Vis. Res.* 32:1493–507
Wichmann FA, Hill NJ. 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept. Psychophys.* 63:1293–313
Wiebel CB, Aguilar G, Maertens M. 2017. Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches. *J. Vis.* 17:1
Wood SN. 2017. *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: CRC Press
Yang JN, Szeverenyi NM, Ts'o D. 2007. Neural resources associated with perceptual judgment across sensory modalities. *Cereb. Cortex* 18:38–45

Annual Review of Vision Science

Volume 6, 2020

Contents

Fifty Years Exploring the Visual System
Joel Pokorny and Vivianne C. Smith 1

Genetic and Environmental Risk Factors for Keratoconus
Sionne E.M. Lucas and Kathryn P. Burdon 25

Minimally Invasive Glaucoma Surgery: A Critical Appraisal of the Literature
David J. Mathew and Yvonne M. Buys 47

Human Organoids for the Study of Retinal Development and Disease
Claire M. Bell, Donald J. Zack, and Cynthia A. Berlinicke 91

Cellular-Scale Imaging of Transparent Retinal Structures and Processes Using Adaptive Optics Optical Coherence Tomography
Donald T. Miller and Kazuhiro Kurokawa 115

Microglia Activation and Inflammation During the Death of Mammalian Photoreceptors
Sarah J. Karlen, Eric B. Miller, and Marie E. Burns 149

Reprogramming Müller Glia to Regenerate Retinal Neurons
Manuela Lahne, Mikiko Nagashima, David R. Hyde, and Peter F. Hitchcock 171

Axon Regeneration in the Mammalian Optic Nerve
Philip R. Williams, Larry I. Benowitz, Jeffrey L. Goldberg, and Zhigang He 195

Retinal Ganglion Cell Axon Wiring Establishing the Binocular Circuit
Carol Mason and Nefeli Slavi 215

Topographic Variations in Retinal Encoding of Visual Space
Alina Sophie Heukamp, Rebekah Anne Warwick, and Michal Rivlin-Etzion 237

Organization, Function, and Development of the Mouse Retinogeniculate Synapse
Liang Liang and Chinfei Chen 261

Signals Related to Color in the Early Visual Cortex
Gregory D. Horwitz 287

Role of Feedback Connections in Central Visual Processing
Farran Briggs 313

Linking Neuronal Direction Selectivity to Perceptual Decisions About Visual Motion
Tatiana Pasternak and Duje Tadin 335

Visual Functions of Primate Area V4
Anitha Pasupathy, Dina V. Popovkina, and Taekjun Kim 363

Coding Perceptual Decisions: From Single Units to Emergent Signaling Properties in Cortical Circuits
Kristine Krug 387

Anatomy and Function of the Primate Entorhinal Cortex
Aaron D. Garcia and Elizabeth A. Buffalo 411

Tuning the Senses: How the Pupil Shapes Vision at the Earliest Stage
Sebastiaan Mathôt 433

Can We See with Melanopsin?
Robert J. Lucas, Annette E. Allen, Nina Milosavljevic, Riccardo Storchi, and Tom Woelders 453

Retinal Image Formation and Sampling in a Three-Dimensional World
Larry N. Thibos 469

Image-Computable Ideal Observers for Tasks with Natural Stimuli
Johannes Burge 491

Measuring and Modeling Visual Appearance
Laurence T. Maloney and Kenneth Knoblauch 519

Visual Search: How Do We Find What We Are Looking For?
Jeremy M. Wolfe 539

Rethinking Space: A Review of Perception, Attention, and Memory in Scene Processing
Monica S. Castelhano and Karolina Krzyś 563

Errata

An online log of corrections to *Annual Review of Vision Science* articles may be found at http://www.annualreviews.org/errata/vision