Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches

Christiane B. Wiebel*
Modeling of Cognitive Processes, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany

Guillermo Aguilar*
Modeling of Cognitive Processes, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin and Bernstein Center for Computational Neuroscience, Berlin, Germany

Marianne Maertens
Modeling of Cognitive Processes, Department of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, Germany

One central problem in perception research is to understand how internal experiences are linked to physical variables. Most commonly, this relationship is measured using the method of adjustment, but this has two shortcomings: The perceptual scales that relate physical and perceptual variables are not measured directly, and the method often requires perceptual comparisons between viewing conditions. To overcome these problems, we measured perceptual scales of surface lightness using maximum likelihood difference scaling, asking observers only to compare the lightness of surfaces presented in the same context. Observers were lightness constant, and the perceptual scales qualitatively and quantitatively predicted perceptual matches obtained in a conventional adjustment experiment. Additionally, we show that a contrast-based model of lightness perception predicted 98% of the variance in the scaling and 88% in the matching data. We suggest that the predictive power was higher for scales because they are closer to the true variables of interest.

Introduction

One major objective in the scientific study of perception is to understand how psychological experiences are linked to physical variables in the world (Fechner, 1860). Devising proper methods to quantify this relationship has turned out to be challenging because psychological variables, contrary to physical ones, cannot be observed directly and must be inferred from observers' responses to properly chosen stimuli (e.g., Gescheider, 1988). In the absence of a well-established measurement theory (Krantz, Luce, Suppes, & Tversky, 1971), Fechner's (1860) simple method of adjustment (matching) is hard to beat and remains widely used (Koenderink, 2013).

To illustrate the problem, let's say we are interested in the perceived lightness of the target check (Figure 1A, red outline) presented behind a transparent medium. Introducing a transparent medium between a surface and the observer (Figure 1B) changes the mapping between surface reflectance and retinal luminance in a characteristic way (Figure 1C). The luminance range of surfaces seen through a transparent medium is substantially reduced and potentially shifted relative to the luminance range for surfaces seen in plain view. To be invariant against such changes, the visual system has to "undo" these changes by appropriate computations (e.g., Singh, 2004; Singh & Anderson, 2002; Wiebel, Singh, & Maertens, 2016). This approximate invariance of perceived lightness across varying luminance is known as lightness constancy. We know from experience and from

Citation: Wiebel, C. B., Aguilar, G., & Maertens, M. (2017). Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches. *Journal of Vision*, *17*(4):1, 1–14, doi:10.1167/17.4.1.

doi: 10.1167/17.4.1 Received November 21, 2016; published April 3, 2017 ISSN 1534-7362

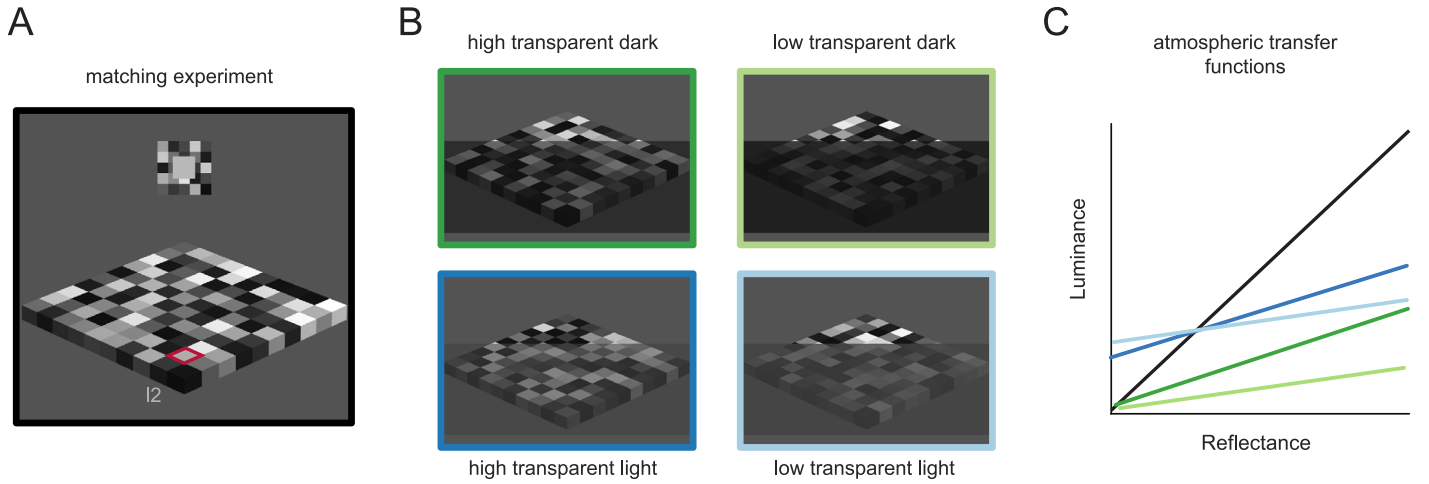

Figure 1. Experimental stimuli. (A) The basic stimulus is a 10 × 10 checkerboard composed of checks with 13 possible reflectance values. In an asymmetric matching task, observers adjust the luminance of an external test field so that it matches the perceived lightness of a specified target check (here *I2*). Observers are said to be lightness constant when their matches indicate the inversion of the various reflectance-to-luminance mappings that are introduced by different transparent media (see panels B and C). (B) Checkerboards were also presented behind different transparent media that varied in reflectance (dark and light) and in transmittance (high and low). (C) ATFs relate target reflectance (*x*-axis) to target luminance (*y*-axis) (Adelson, 2000). The color scheme corresponds to the images in panel B. In the transparency conditions, the luminance range is compressed and/or shifted with respect to plain view. This is reflected in corresponding slope and intercept changes of the respective ATFs.

empirical studies that human observers are indeed largely invariant against such fluctuations in retinal luminance. However, we still lack a theoretical model of how the visual system accomplishes lightness constancy. To develop such a model, we must be able to measure the relationship between retinal luminance and perceived lightness in a reliable and comprehensive way. To that end, we ideally want to estimate the functions describing this relationship, which are known as transducer functions or perceptual scales (e.g., Kingdom & Prins, 2010).

The most commonly used method for measuring this relationship is the method of adjustment even though it does not provide a direct estimate of the transducer functions and it presumes a number of operations on the part of the observer. Figure 2 depicts the processes involved in adjustment or matching procedures for perceived lightness. An observer adjusts the intensity of a test stimulus so that it looks identical to a given standard. It is assumed that the observer internally compares magnitudes of perceived lightness for the target ($\Psi[x_T]$) and the match ($\Psi[x_M]$). What is being

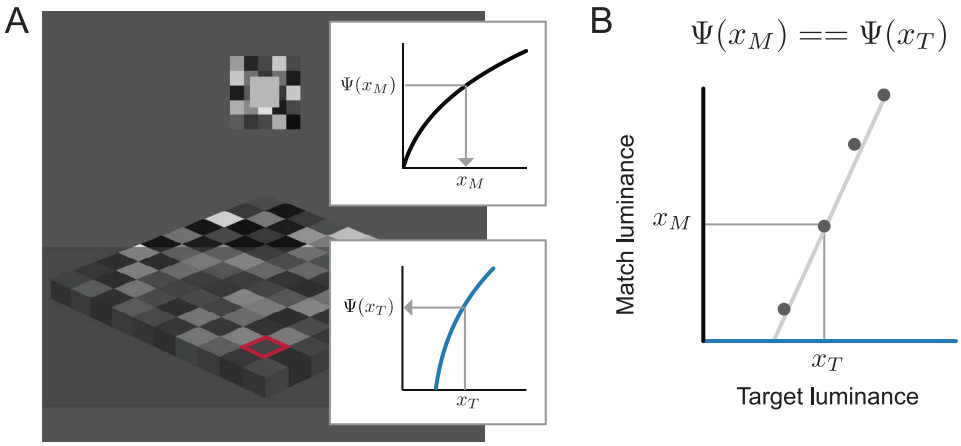

Figure 2. Perceptual processes underlying matching procedures. (A) At each position, match and target, there is a transducer function that relates retinal luminance (x_M, x_T) to perceived lightness [$\Psi(x_M)$, $\Psi(x_T)$, insets on the stimulus]. (B) What is measured in a matching procedure are the luminances x_M and x_T that correspond to equal perceived lightness at both positions [$\Psi(x_M) == \Psi(x_T)$]. After Maertens and Shapley (2013).

measured, however, are not the transducer functions relating the two but the corresponding luminances of the target and the match (x_T and x_M, Figure 2B).

Another problem with the method arises when, as in the above case, test and match are presented in different contexts (asymmetric matching). In most cases, researchers are interested in such asymmetric comparisons because they allow one to quantify the degree of perceptual constancy. Such asymmetric comparisons become problematic, however, when the difference in context causes appearance differences that cannot be compensated along the dimension of the adjustment (Brainard, Brunt, & Speigle, 1997; Ekroll & Faul, 2013; Foster, 2003; Logvinenko & Maloney, 2006; Logvinenko, Petrini, & Maloney, 2008). The consequence would be an inaccurate or even invalid measurement that does not capture the perceptual representation of the stimulus.

Recently, there have been attempts to tackle the problems associated with matching (Logvinenko & Maloney, 2006; Logvinenko et al., 2008; Radonjić & Brainard, 2016; Radonjić, Cottaris, & Brainard, 2015b; Umbach, 2013). Although it is widely accepted that observers are relatively lightness constant under natural viewing conditions, many experiments still find varying amounts of constancy for different viewing conditions, stimuli, task types, or even instructions (Foster, 2011; Gilchrist et al., 1999). Such deviations might either be a consequence of methodological problems such as the ones just outlined or a meaningful deviation from constancy, which then would need to be explained by any successful lightness model. Progress in revealing the underlying mechanisms for lightness perception is therefore tightly coupled with choosing appropriate and robust experimental methods that allow the comprehensive testing of theoretical models.

As an effort in this direction, we address the limitations of matching procedures by adopting the following approach. We measure the transducer functions directly using maximum likelihood difference scaling (MLDS, Maloney & Yang, 2003). MLDS is a scaling method that allows the efficient estimation of perceptual scales, i.e., the transducer functions relating retinal luminance and perceived lightness (Figure 1). It has been used to study various perceptual dimensions (e.g., Fleming, Jäkel, & Maloney, 2011; Obein, Knoblauch, & Viénot, 2004). Furthermore, it is based on a signal detection model, which potentially allows one to relate measurements of appearance with measurements of discriminability (Aguilar, Wichmann, & Maertens, 2017; Devinck & Knoblauch, 2012). Here we used MLDS to measure perceptual scales in different contexts using only within-context comparisons in order to avoid the procedural problems of asymmetric matching. The estimated scales are constructed from the judgment of perceived stimulus differences and not from the adjustment of a reference as in other scaling methods, such as magnitude estimation (Gescheider, 1988) or partition scaling (Whittle, 1994). MLDS requires a straightforward perceptual judgment and is thus less susceptible to strategic influences.

To scrutinize whether MLDS provides reliable perceptual scales of lightness, we validate the scales empirically and theoretically. First, we use the estimated scales to predict perceptual matches and compare them to matches gathered in an independent asymmetric matching experiment. Second, we compare the predictive power of a contrast-based lightness model (Wiebel et al., 2016; Zeiner & Maertens, 2014) for scaling and matching data. To anticipate, we found that (a) the empirical perceptual scales for different contexts were consistent with lightness constancy, (b) matching data were well predicted by the perceptual scales, and (c) human lightness perception followed a difference scale that corresponds to a normalized contrast metric. The predictive power of the contrast-based lightness model was higher for the scaling than for the matching data, suggesting that estimating perceptual scales has the advantage of probing more directly the internal dimension under study.

Methods

Observers

Ten naive observers participated in the study; five of them were female. Observers' ages ranged from 19 to 32 years. All observers had normal or corrected-to-normal visual ability and were reimbursed for participation. Informed written consent was given by all observers prior to the experiment.

Stimuli and apparatus

Stimuli were presented on a linearized 21-in. Siemens SMM2106LS monitor (400 × 300 mm, 1024 × 768 px, 130 Hz). Presentation was controlled by a DataPixx toolbox (Vpixx Technologies, Inc., Saint-Bruno, QC, Canada) and custom presentation software (http://github.com/TUBvision/hrl). Observers were seated 110 cm away from the screen in a dark experimental cabin. Observers' responses were registered with a ResponsePixx button-box (VPixxTechnologies).

The stimuli were images of customized checkerboards composed of 10 × 10 checks (Figure 1). The images were rendered using Povray (Persistence of Vision Raytracer Pty. Ltd., Williamstown, Victoria, Australia, 2004). The position of the checkerboard, the

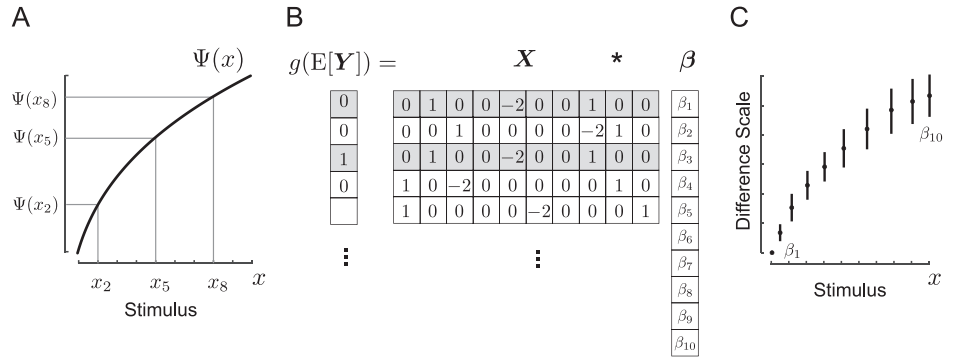

Figure 3. Estimation of scales using MLDS with triads. (A) Hypothetical scale relating perceptual experiences $\Psi(x_i)$ to stimulus values x_i. For an example stimulus triad (x_2, x_5, x_8), observers are asked to compare which pair is more different, (x_2, x_5) or (x_5, x_8). The decision model for this example triad is $\Delta = [\Psi(x_8) - \Psi(x_5)] - [\Psi(x_5) - \Psi(x_2)]$ or $\Delta = \Psi(x8) - 2\Psi(x_5) + \Psi\ (x_2)$. (B) The weights for each term in the decision model are the covariates in a design matrix, *X*, in a binomial GLM. Each row in the design matrix *X* indicates the elements of a triad in one trial. The shaded rows are two repetitions of the same triad, which is the example triad in panel A. The design matrix contains all triads shown in one experiment. (C) Empirical scale that results from solving the GLM and obtaining the coefficients β. The coefficients correspond to the scale values at different levels of the physical variable. Error bars are estimates for the error associated with each coefficient and were obtained using bootstrap.

light source, and the camera were kept constant across all images. Checks were assigned one out of 13 surface reflectance values according to the experimental design (see below). In the transparency conditions, a transparent layer was placed between the checkerboard and the camera (Figure 1B). It was positioned so as to cover all targets and their surrounding checks in both the MLDS and the matching experiment. The transparency was created using alpha blending (Metelli's episcotister model). The image luminances of the background B and the foreground F are combined according to some weighting factor α so as to result in a new image luminance at the position of transparency $T = \alpha \times B + (1 - \alpha) \times F$. An α value of 0 corresponds to an opaque foreground $T = F$; α of 1 corresponds to a fully transparent foreground $T = B$. The transparent layer varied in transmittance and reflectance. The dark transparency had a value of 0.35 in *povray* reflectance units (19 cd/m^2) and the light transparency of 2 (110 cd/m^2). Values of $\alpha = 0.4$ and 0.2 were used in the high and low transmittance conditions, respectively. The rendered images were converted to grayscale images. The background luminance was 141 cd/m^2. Detailed values of luminance for each transparent medium can be found in Supplementary Table S3).

In the matching experiment, an adjustable test field was presented above the checkerboard to assess observers' lightness matches (Figure 1A). The test field was embedded in a coplanar surround checkerboard that was composed of 5×5 checks. The size of the test field was $1.2° \times 1.2°$ visual angle and that of the surround checkerboard was $3° \times 3°$. The luminances of the checks in the surround checkerboard were fixed throughout the experiment, and the luminances were chosen so that two adjacent checks did not have the same luminance. The mean luminance of the surround checks was 178 cd/m^2, which is identical to the mean luminance of the 13 checks in the main checkerboard in plain view. The surround checkerboard was presented in four different spatial arrangements, resulting from clockwise rotation of the original in steps of 90°. A configuration was assigned randomly to each trial.

Design and procedure

Perceptual scales and asymmetric matching functions were measured for five different viewing conditions, a plain view condition, and four transparency conditions (Figure 1).

MLDS experiment

We used MLDS with the methods of triads (Figure 4A; Knoblauch & Maloney, 2008, 2012). We used 10 out of the 13 reflectance values to construct the triads. The lowest and the two highest reflectance values were omitted to achieve a feasible number of trials. With $p = 10$ reflectance values, the total number of unique triads was $n = p!/((p-3)! \times 3!) = 10!/(7! \times 3!) = 120$. Each triad contained three values that were selected so as to enclose nonoverlapping intervals. They were presented in ascending ($x_1 < x_2 < x_3$) or descending ($x_1 > x_2 > x_3$) order (Knoblauch & Maloney, 2008). The reference, x_2 (check *I2* in Figure 4A), was located between the two comparisons, x_1 and x_3 (checks *B2* and *I9* in Figure 4A). In each trial, observers judged which comparison check, x_1 or x_3, was more different in lightness from the reference. Observers used a left or right response button to indicate their choice. No time limit was imposed.

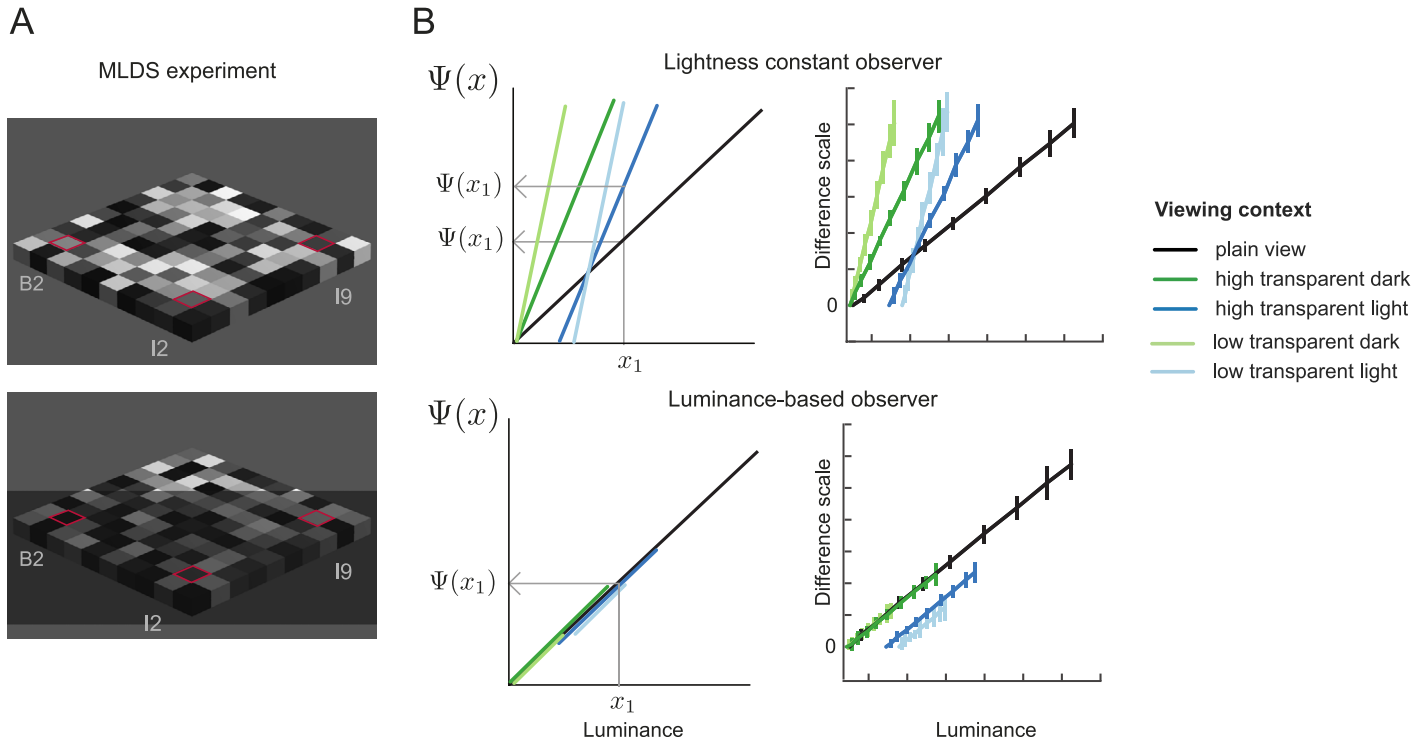

Figure 4. Method of triad procedure and observer models. (A) In the triad comparison, observers compared the lightness of three specified checks (*B2*, *I2*, and *I9*, marked with a red outline). The upper panel shows a triad comparison in plain view, the lower panel a comparison behind one of the transparent media. (B) Simulation for a lightness-constant (upper panels) and luminance-based observers (lower panels). For the lightness-constant observer, the perceptual scales (upper left panel) correspond to an inverse mapping of the ATFs (Figure 1C). For the luminance-based observer (lower left panel) the luminance-to-lightness mappings in different contexts coincide on a single function. We generated data for each of the models in simulations, and the estimated perceptual scales are shown on the right panels. See text for details.

To keep the local context comparable for the elements of a triad, we controlled the luminances of the eight checks surrounding each triad element. The same eight luminance values were used for each triad element, but they differed in spatial arrangement. Their mean luminance was 178 cd/m^2, which was identical to the mean luminance of all checks seen in plain view. The remaining 73 checks were drawn randomly without replacement from a set consisting of six repeats of the 13 different reflectance values. This resulted in a slight variation of the mean luminance of those checks between trials (up to 6 cd/m^2). The checks were positioned so that two neighboring checks did not have the same reflectance.

Each triad was repeated 10 times, resulting in 1,200 trials per viewing condition and 6,000 trials in total. Trials were randomized across viewing condition, triad, and target reflectance. The experiment was divided into several sessions. A new image was created for each trial.

Matching experiment

Target reflectances and viewing conditions were identical to those in the MLDS experiment. The target check was presented at the position of the reference (check *I2* in Figure 1) in the MLDS experiment. Observers adjusted the luminance of the external test field to match the perceived lightness of the target check. The luminance was adjusted by pressing one of four buttons, two of them for coarse adjustments (±10 cd/m^2) and the other two for fine adjustments (±1 cd/m^2). The maximum luminance of the monitor was 550 cd/m^2. Satisfactory matches were confirmed with a fifth button that initiated the next trial. No time limit was imposed on the adjustment procedure.

The eight checks surrounding the target were assigned in the same way as in the MLDS experiment. The remaining 91 check reflectances were drawn randomly without replacement from a set consisting of eight repeats of all 13 reflectance values. Thus, the mean luminance across trials was comparable to that in the MLDS experiment. Again, neighboring checks had to have different reflectances.

Each combination of target reflectance and viewing condition was repeated 10 times, resulting in a total of 500 trials. A new image was created for each trial, and trials were randomized across experimental conditions.

MLDS analysis

In the following, we explain how we used the MLDS routines (Knoblauch & Maloney, 2008) in *R* to analyze the data. The left panel in Figure 3 depicts a hypothetical perceptual scale that relates psychological experience, $\Psi(x)$, to a physical variable, x. The central panel illustrates how the decision model translates into the statistical model that is used to estimate scale parameters, and the right panel depicts the estimated scale values. Observers perform the triad judgments for different levels (x_i) of the physical variable (e.g., x_2, x_5, x_8 in Figure 3). They judge whether the difference $\Delta = (\Psi[x_8] - \Psi[x_5]) - (\Psi[x_5] - \Psi[x_2])$ is smaller or larger than zero. The decision model for all possible triads is summarized in the design matrix X, which contains separate columns for each x-value (Figure 3B). Each row of the design matrix contains the weights for the decision model of that respective triad. The coefficients (β) are estimated in a (binomial) generalized linear model (GLM) to account for the observed responses (Y) using maximum likelihood, and they represent the scale values for all levels of the physical variable. The linear predictors $X * \beta$ are related to the observed responses by using a link function $g()$, which maps the range of the linear predictors to a range of the response probabilities $E[Y]$. The decision model in MLDS is stochastic, and it assumes a single Gaussian-distributed noise source ε that corrupts the decision variable. By default, the GLM estimation assumes a variance of the noise source of one ($\sigma^2 = \sigma_\Delta^2 = 1$), and as a consequence, the amount of noise estimated by the model is inversely related to the scale's maximum, with a higher maximum when the estimated noise is low (so-called "unconstrained" scales in Knoblauch & Maloney, 2008). However, alternative parameterizations are also possible. Within the framework of the GLM, the scaling can be controlled fairly simply by prescaling the design matrix. For example, dividing the weights in the matrix X in Figure 3 by two—giving 0.5, −1, and 0.5—yields a scale for which $\sigma_\Delta^2 = 4$ that corresponds to $\sigma_\beta^2 = 1$ for each lightness level. This would parameterize the scale in terms of d' (as shown in more detail in Aguilar et al., 2017; Devinck & Knoblauch, 2012).

Simulation of observer models

We used an ideal observer analysis to test whether MLDS could distinguish between different generative models. In particular, we tested a lightness constant against a luminance-based observer, two extremes of behavioral judgments. The model comparison is done as follows. We define internal scales for each of the two models (Figure 4B). For a luminance-based observer, the luminance-to-lightness mappings in different contexts coincide on a single function and differ only in the range of luminance values (Figure 4B, lower left panel). Formally, the sensory representation function was defined as

$$\Psi^{lum}(x) = a \cdot x + b,$$

where x is luminance, and a, b are linear coefficients calculated to map the range of luminance in plain view [$L_{\min}$, $L_{\max}$] to the range [0, 1].

For a lightness-constant observer, the mapping functions in different contexts should "undo" the transformations of image formation in which equal surface reflectances are mapped onto different luminance ranges (Figure 1C). Thus, we model this observer by using internal mapping functions that are the inverse functions of the atmospheric transfer functions (ATFs) shown in Figure 1C. Formally, the sensory representation function was defined as

$$\Psi^{\text{light}}(x) = a_i \cdot x + b_i \qquad i \in 1...5,$$

where x is luminance, and a_i, b_i are linear coefficients calculated to map the range of luminance for each viewing condition to the range [0, 1] (for simplicity, we used linear functions, but power functions could be used as well and would not change our ideal observer results).

Each of the two observer models is used to generate responses in a "mock" MLDS experiment that has the same number of triads and repetitions as the actual experiment. For each triad and repetition, the decision variable was calculated as

$$\Delta = [\Psi^*(x_3) - \Psi^*(x_2)] - [\Psi^*(x_2) - \Psi^*(x_1)] + \epsilon \tag{1}$$

with $\epsilon \sim N(0, \sigma^2)$, and Ψ^* is either Ψ^{lum} or Ψ^{light}. Simulated responses were generated choosing the triad (x_2, x_3) when $\Delta > 0$ and (x_1, x_2) otherwise. Finally, the simulated data were subjected to the MLDS analysis to obtain the coefficients β that constitute the scale values. Figure 4B shows the model perceptual scales (left) and the estimated scales (right), and it is evident that for the chosen noise level ($\sigma = 0.15$) the method recovers the underlying scale.

We repeated the ideal observer analysis for a range of different noise levels (σ, minimum = 0.01 and maximum = 1.2, see Supplementary Material). The two observer models were distinguishable for a broad range of noise levels up to approximately 0.4. This upper-bound value was much higher than the noise levels that have been observed in previous experiments (Devinck & Knoblauch, 2012; Knoblauch & Maloney, 2008). We therefore concluded that MLDS could be used to derive meaningful scales because they would allow us to distinguish between these two different observer models.

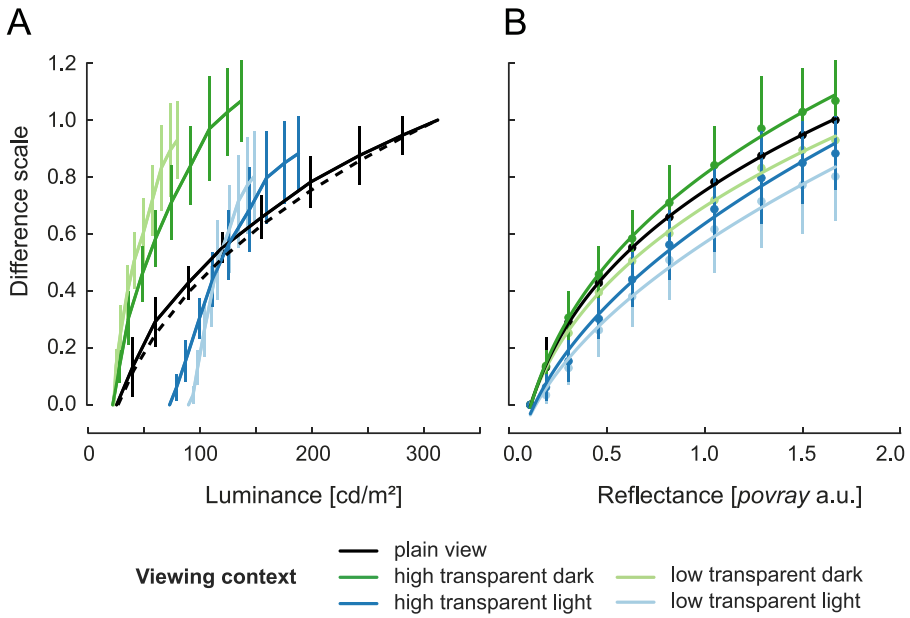

Figure 5. MLDS difference scales in different viewing conditions. (A) Difference scales as a function of luminance. The functions depict the aggregated scales across observers ($n = 10$). For each observer, the scales were normalized with respect to plain view and then aggregated. The dashed black line depicts the Munsell scale in plain view (see main text for a description of the Munsell scale). Error bars indicate $M \pm SD$. (B) Same as in panel A, but scales are plotted as a function of reflectance. Scale values (markers, $M \pm SD$) were fitted with a power function (lines) individually for each viewing condition.

Results

Figure 5 shows the perceptual scales measured in different viewing conditions aggregated across all observers. The scales are interval scales with the minimum anchored at zero and the maximum being inversely proportional to the estimated noise (in MLDS terminology referred to as "unconstrained scales"; Knoblauch & Maloney, 2012).

The empirical scales are consistent with a lightness-constant observer and not with a luminance-based observer. This is evident from a comparison between the model predictions (Figure 4B) and the observed result pattern (Figure 5A). Although the estimated scales are not linear, they share crucial features with the hypothetical scales. First, there is a difference in "intercept" between perceptual scales in the light and dark transparency conditions (blue vs. green lines in Figure 5A). Second, the scales are steeper for transparent media with lower transmittance than with higher transmittance (light vs. dark colored lines in Figure 5A). Figure 5A also plots the Munsell neutral value scale (Munsell, Sloan, & Godlove, 1933) that would be predicted for our choice of luminances (dashed black line in Figure 5A).

The Munsell scale represents the expected scale that relates equal steps in perceived lightness to luminance (Whittle, 1994). It was calculated by setting the highest luminance in the plain view stimulus as the white reference, i.e., to the maximum of one (Pauli, 1976). It is evident from Figure 5A that the Munsell scale is consistent with the perceptual scale estimated in our plain view condition. The typical nonlinear shape indicates higher sensitivity for differences between checks with low reflectances than for checks with high reflectances. This has indeed been reported in previous work (e.g., Chubb, Landy, & Econopouly, 2004). To aggregate scales across observers, we normalized the scales of each individual observer relative to the maximum scale value in the plain view condition. The ranges of the scales differed between observers because different observers have different noise levels. The data for individual observers are provided in Supplementary Figure S1.

Scales as a function of reflectance

To better illustrate the degree of lightness constancy across conditions, we replaced luminance by reflectance at the *x*-axis of the perceptual scales. In such a perceived lightness versus reflectance plot, the scales of a lightness-constant observer should coincide on a single function. Figure 5B shows that this was indeed that case.

To assess the agreement between scales in different conditions quantitatively, we compared the functions that were fit in each condition against what we call a global fit in which the data from all conditions are fitted

by a single function. If the data in different viewing conditions can be explained by one internal model, then the global fit should account for the data as well as the individual fits for each viewing condition. We fitted the scale parameters in each condition and the global scale with a power function $\Psi(x) = ax^e + b$ using a nonlinear least squares method (Ritz & Streibig, 2008). To evaluate the goodness of fit, we computed R^2 values for linear fits to the data. The average R^2 was already reasonably high (0.86). We then performed F tests on nested models (power function vs. its linear submodel with $e = 1$), which revealed that the power functions fitted the data significantly better than the linear ones, $F_{\min}(1, 97) = 15.6$, $p < 0.001$. From this, we concluded that the power functions captured the data sufficiently well.

We used a GLM to test whether applying single models to the data in the five different viewing conditions would result in better fits than applying a global model to all data. We compared the respective sum of squares for the global model with three parameters (a, b, e) and for the separate models with 5×3 parameters. There was a benefit for the separate model fits relative to the global model, $F(12, 497) = 18.57$, $p < 0.001$. To explore the cause for this difference, we computed one-way repeated-measures ANOVAs for each of the three parameters of the power functions. We found a significant difference between scales for the exponent parameter, e, $F(4, 36) = 16.6$, $p < 0.001$, which determines the curvature of the function. Post hoc tests on the exponents revealed significant differences between each of the light transparency conditions and the plain view and the dark transparency with high transmittance (Bonferroni corrected $p < 0.05$). The main difference between the light transparency conditions and the plain view and the dark transparency (high transmittance) conditions is the difference in curvature between these functions (Figure 5B).

The light transparency conditions are special insofar as during image formation the reflectance-to-luminance mapping undergoes a range reduction and a range shift (see Figure 1). This means that checks seen through a light transparent medium undergo the greatest compression in its contrast range. The Michelson contrast for targets in plain view range from −0.84 to 0.4 whereas in the low transparent light condition they range from −0.16 to 0.16 (the contrast is computed relative to the mean luminance in the region of transparency). Therefore, sensitivity might be lower for this range of the stimuli.

Perceptual scales and matching functions

We illustrated in Figure 2 how the data recorded in matching procedures are related to perceptual scales. Here, we show to what extent the theoretical relationship can be corroborated by experimental data. To predict matching data from perceptual scales, we needed to first find the scale value $\Psi(x_T)$ that corresponds to a particular target luminance x_T in one of the transparency conditions. In the next step, we needed to find the luminance value x_M that corresponds to the scale value at the match position $\Psi(x_M)$, assuming that observers match the lightness of the match region to that of the target region according to $\Psi(x_M) = \Psi(x_T)$. We did not measure a perceptual scale at the match position but instead adopt the plain view scale to represent the scale for the matches. In order to be able to read out x-values corresponding to any possible Ψ-value and vice versa, we fitted the scales with power functions, $\psi(x) = ax^e + b$, using a nonlinear least squares method. We derived the predicted matching data from the "unconstrained" scales individually for each observer, and we then aggregated them in the same way as the empirical data obtained from the matching experiment.

In Figure 6, empirical and predicted matches are plotted next to each other (panels A and B, respectively), and it can be seen that they share some characteristic features. The matching functions, like the scales (Figure 5A), differ in slope and intercept between the different transparency conditions. Differences in transmittance are accompanied by differences in slope, and differences in reflectance are accompanied by differences in intercept. Unlike the scales, the matching functions are linear.

For a quantitative evaluation of the degree of similarity between empirical and predicted matching data, we computed linear regressions for each of the viewing conditions. We used within-subject t tests to compare slopes and intercepts between predicted and empirical functions. The average slope and intercept values are listed in Supplementary Table S4 together with the relevant test statistics. We found significant differences between the predicted and the empirical functions only for the dark transparent medium with a high transmittance.

Predictive power of a contrast-based model

The estimated perceptual scales are an interesting test case for lightness models because they represent a more direct measurement of perceived lightness than the matching data. In particular we compared how well our previously suggested normalized-contrast model (Zeiner & Maertens, 2014) could account for both the scaling and the matching data.

The normalized contrast model was initially motivated by the observation that the introduction of a transparent medium leads to a systematic change in contrast range of the respective image region. It was

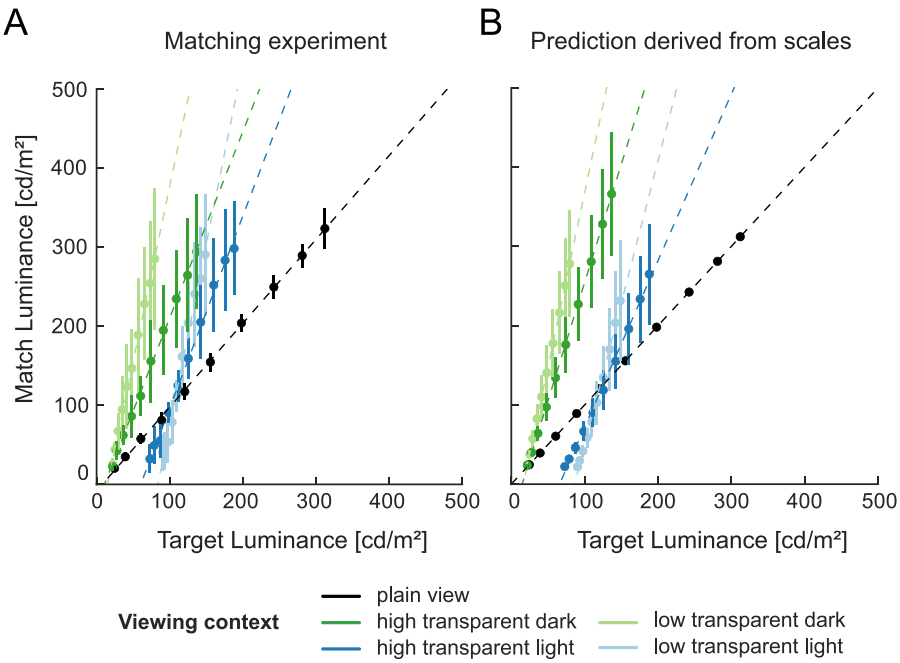

Figure 6. Empirical and predicted matching data. (A) Results of the matching experiment. The luminance adjusted in the matching field (*y*-axis) is plotted as a function of target luminance (*x*-axis) in each viewing condition. Data were aggregated across observers ($n = 10$). Error bars indicate $M \pm SD$. (B) Same as in panel A but for matches predicted from the estimated MLDS scales.

suggested that this change in contrast range might serve as a cue to segregate the region from regions seen in plain view (Anderson, 1999; Singh, 2004; Singh & Anderson, 2002). It has been subsequently shown that the accompanying contrast statistics can be used to accurately predict perceived lightness (Singh, 2004; Singh & Anderson, 2002; Wiebel et al., 2016; Zeiner & Maertens, 2014). The normalized contrast model engages two processing steps: First, the target intensity is normalized relative to its local surround by computing the Michelson contrast between target and surround. Second, this target contrast is normalized relative to the contrast range in the region of the transparency, which is subsequently mapped to the contrast range in plain view (for details of the normalized contrast model calculation, see Supplementary material). The so-derived normalized contrast predicts observers' lightness matches in contrast units.

Figure 7 shows the aggregated data of both experiments as a function of the model predictions. If the computed normalized contrast accounts well for differences in appearance, then the functions should line up on top of each other, and they should become more linear (see Knoblauch & Maloney, 2012, for a similar rationale underlying correlation perception). Transforming the *x*-axis into units of normalized Michelson contrast did indeed linearize the perceptual scales. To test how well the normalized contrast model accounts for the variability between the different context conditions, we computed a global R^2 value. As described before, we treat all data as if they were coming from one underlying model. The normalized contrast measure accounts for 98% of the variance in the scaling data and for 88% of the variance in the matching data. This indicates that the normalized contrast measure is a better predictor for the scales than the matching data by explaining more variance. The residuals of these fits are provided in Supplementary Figure S6.

Discussion

The goal of this work was to better understand how psychological experiences are linked to physical variables. We studied the question in the domain of lightness perception, but the observed principles equally apply to other domains of perceptual appearance. To make progress toward that goal, we measured perceptual scales that link perceived lightness to image luminance using MLDS. Our results show that the estimated perceptual scales (a) are consistent with a lightness-constant observer model in all viewing contexts, (b) predict perceptual equality across different viewing contexts, (c) indicate that human lightness perception follows a difference scale that corresponds to a normalized contrast metric. The normalized contrast model accounted for more of the variance in the scaling (98%) than in the matching data (88%), suggesting that estimating perceptual scales has the advantage of probing more directly the internal lightness scale.

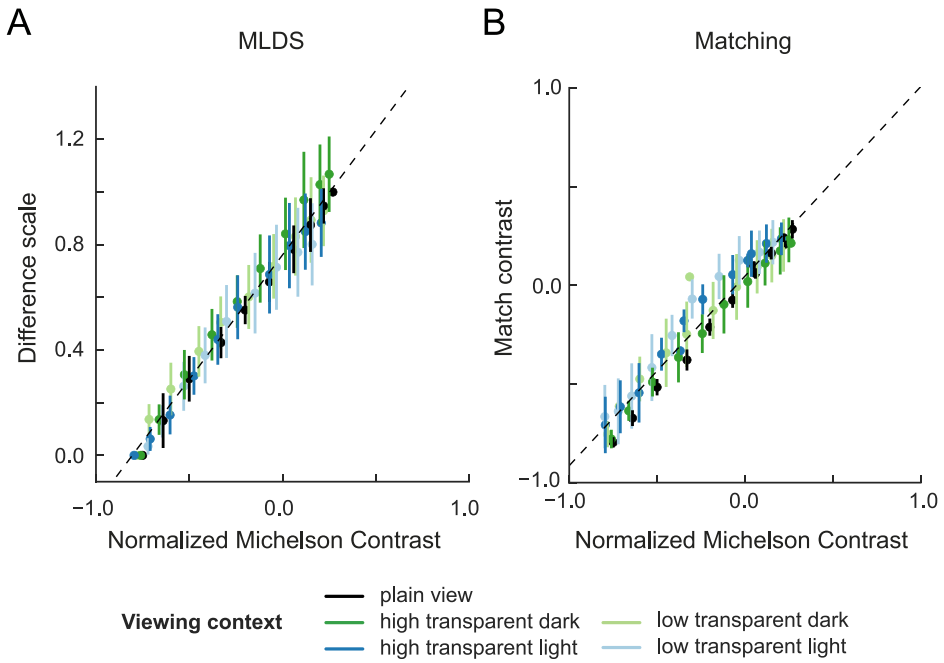

Figure 7. Perceptual scales (Figure 5A) and matching data (Figure 6A) plotted as a function of the normalized Michelson contrast. Dashed lines indicate a linear fit to the data for all viewing contexts ($R^2 = 0.98$ for MLDS, $R^2 = 0.88$ for matching). Error bars indicate $M \pm SD$ across observers.

MLDS-based lightness scales

The estimated perceptual scales were in close correspondence with each other (Figure 5B), i.e., perceived lightness followed the actual check reflectances despite substantial variations in check luminance across viewing conditions. This reflects a high degree of lightness constancy. This was corroborated by the simulated observer models because the empirical scales were consistent with the lightness-constant and not the luminance-based observer. The shape of the perceptual scales followed the shape of the classical Munsell scale. The perceptual scales are an estimation of the transducer functions, which cannot be uncovered using matching (see Figure 2).

In addition to the MLDS experiment, we conducted a conventional asymmetric matching experiment. We tested to what extent the postulated relationship between perceptual scales and matching (Figure 2) would be evident in the data. Predicted and empirical matching functions were consistent with each other (Figure 6). The high degree of consistency is noteworthy because triad comparisons and matching require different perceptual judgments. Asymmetric matching can be likened to measuring a rod of unknown length with a ruler whereas in triads rods of different lengths would be compared among each other. The consistency between both types of measurements indicates that the stimulus suitably constrains the perceptual response to judgments based on lightness and not on luminance. This cannot be taken for granted (Arend & Goldstein, 1987; Radonjić & Brainard, 2016), in particular because observers were not explicitly told what dimension to judge.

A potential challenge when comparing perceptual scales measured in different contexts is the necessary assumption of how scales are anchored. Two perceptual scales might have the same shape but cover a different range, implying a different anchoring. Classical scaling experiments did not confront this problem because perceptual scales were measured in only one context, i.e., plain view. The default of MLDS is to anchor the perceptual scales at zero. This is an arbitrary choice, and any linear transformation of the scale would be a valid outcome of the analysis. The good correspondence between the estimated scales and the matching data in the present case suggests that there was no substantial anchoring problem.

As described by Knoblauch and Maloney (2012), MLDS assumes that observers are stochastic in their judgments with the noise originating at the decision level (as shown in Equation 1). This assumption implies that observers are worse at judging interval differences that are small, i.e., when $(\Psi[x_3] - \Psi[x_2]) \sim (\Psi[x_2] - \Psi[x_1])$. This critical assumption in MLDS is different than other scaling methods, such as Fechnerian scaling that uses integration of just-noticeable differences or other discrimination-based scaling methods (Baird, 1978). These scaling methods assume a noise source at an early sensory representation level and not at a late decision level. Here, we compared perceptual lightness

scales that were measured in different viewing conditions and hence could have been associated with different amounts of decision noise. This was not what we observed. Although individual observers differed in their overall noise level, all scales measured for one observer had comparable estimated noise levels. However, these assumptions must be considered carefully, and ultimately their validity must be addressed experimentally (Aguilar et al., 2017).

The estimated noise level is critical for the interpretation of scales as well as with respect to the distinction between our two observer models (lightness constant vs. luminance-based). This is possible only up to a limit at which observers' noise is too large for the models to be distinguished. We established in simulation that this upper bound is at an estimated noise level $\hat{\sigma} = 0.4$ (Supplementary Material). In our observers the estimated $\hat{\sigma}$ values varied from 0.13 to 0.21 for observers O1 to O8, i.e., values below the upper limit of model discriminability. For observer O9, $\hat{\sigma} = 0.39$ was at the boundary of discriminability, and for observer O10, $\hat{\sigma} = 0.71$ was beyond the upper limit. Thus, the noise level of observer O10 did not allow a definite selection of either of the two models. The estimated noise level also must be considered carefully when comparing scales against ideal observer models.

Alternatives to asymmetric matching

Asymmetric matching has been criticized in the past for two main reasons: First, observers' matches reflect the underlying perceptual magnitudes only indirectly (Gescheider, 1988; Maertens & Shapley, 2013). Second, observers' matches might not reflect perceptual identity but merely the best possible match (Brainard et al., 1997; Ekroll & Faul, 2013; Foster, 2003; Logvinenko & Maloney, 2006). In particular, the question whether lightness is represented by more than one dimension across different contexts has been tackled using different methods (Logvinenko & Maloney, 2006; Logvinenko et al., 2008; Umbach, 2013). Beyond methodological shortcomings, asymmetric matching tasks have also been criticized for their lack of realism because in real life we rarely adjust the color of an object but rather select objects based on their color. In two recent studies, Radonjić, Cottaris, and Brainard (2015a, 2015b) measured color constancy in a color selection paradigm in which they asked observers to select which of two competitors was more similar to a given target.

Their task was analogous to the triad comparison used in MLDS, but the design was different from the standard MLDS design. A limited number of targets was presented as anchor for a respective set of competitors, but these competitors were not compared with each other. MLDS would involve triad comparisons of all possible combinations of targets and competitors.

The data were analyzed with a customized version of MLDS. The crucial difference from our approach is that in their critical condition target and competitors were presented in different illuminations. As a consequence, observers' judgments were subject to the same comparison problem as in asymmetric matching. To estimate a perceptual scale, it was assumed that target and competitors are represented on a common underlying dimension. In our way of thinking, this means to skip the step of estimating the different transducer functions (scales), which map luminance to perceived lightness in different contexts (Figure 4B), and to compare stimuli directly on the internal axis. As we have outlined above, this assumption is valid only for a lightness constant observer, i.e., for observers whose perceptual scales in different viewing situations have comparable scale maxima. The authors reported moderately high color constancy indices, which were comparable to asymmetric matches for the same type of stimuli (Radonjić et al., 2015a, 2015b). We suggest including such cross-context comparisons only to validate predictions from the MLDS-based scales as we did here with the asymmetric matches.

Models of lightness perception

We claim that perceptual scales are an important test case for models of lightness perception because they offer a direct estimate of the transducer functions that we are interested in. A successful model should be able to explain both characteristics of lightness appearance: perceptual equality across contexts as well as sensitivity differences manifested in the shape of perceptual scales (Hillis & Brainard, 2007).

If we assume that the goal of the visual system is to accurately represent surface reflectance, then reflectance would be the best predictor of perceived surface lightness. Thus, for a perfect lightness-constant observer, the perceptual scales measured in different contexts should perfectly overlap when plotted against reflectance. Our empirical scales are consistent with a lightness-constant observer; however, they reveal small deviations, especially for the two lighter transparent media (Figure 5B). When we plotted the scales as a function of normalized contrast (Wiebel et al., 2016; Zeiner & Maertens, 2014) instead of reflectance, the differences between scales were substantially reduced (Figure 7A). This means that the normalized contrast metric does not perfectly capture *veridical* surface reflectances but is rather tightly correlated with them. One might be tempted to conclude that the predictive

power of the contrast-based model "exceeds" that of physical surface reflectances because it accounts for the deviations from lightness constancy that we observed in the data.

This finding is consistent with the idea that the visual system, instead of doing inverse optics (e.g., Barrow & Tenenbaum, 1978; D'Zmura & Iverson, 1993), might use a set of readily available but imperfect cues to infer stable properties of objects (e.g., Anderson, 2011; Fleming, 2014). The involved computations might not always lead to a *veridical* percept with respect to the physical world, but to an overall reliable estimate of the appearance of objects (e.g., Marlow, Kim, & Anderson, 2012). The estimated scales were linearized by the transformation to contrast units, which implies that the model accounts for the sensitivity differences between low and high reflectances (e.g., Lu & Sperling, 2012), a feature that cannot be quantitatively captured with matching. The higher agreement between the model and the perceptual scales (compared to matching) supports the idea that the perceptual scales are a more direct and informative measure of the internal variable of lightness and subject to fewer sources of variability.

General conclusions

In this paper, we show that a scaling method is more powerful than matching in elucidating the perceptual representation of surface lightness. MLDS provides a direct estimate of the transducer functions that relate the physical dimension of reflectance to the psychological dimension of perceived lightness. In addition, MLDS avoids the practical difficulties associated with asymmetric matching tasks because all perceptual comparisons are made within the same viewing context. Observers confirmed that subjectively the triad comparison required by MLDS was a natural and straightforward task.

So why is it then that asymmetric matching remains the method of choice despite the obvious benefits of MLDS. We suspect that experimenters feel slightly uneasy about explicitly making and committing to the various assumptions that are required by MLDS in order to statistically estimate the perceptual scales. However, as we illustrate in Figure 2, asymmetric matching procedures also assume the presence of internal scales, but they are hidden, and their shape cannot be inferred from observers' matches. We think that the present results are encouraging and advocate the estimation of scales because they provide a more direct estimate of internal variables against which we can test our theoretical models of appearance.

Keywords: lightness, perceptual scales, MLDS, asymmetric matching

Acknowledgments

This work has been supported by an Emmy-Noether research grant of the German Research Foundation (DFG MA5127/1-1) and by the Research Training Grant "Sensory Computation in Neural Systems" of the German Research Foundation (GRK1589/1-2). We would like to thank Michael Landy, Kenneth Knoblauch, Bart Anderson, Richard Murray, Felix Wichmann, Frank Jäkel, and David Higgins for their constructive suggestions that helped in improving this manuscript. CBW has moved in the meantime to the Honda Research Institute Europe in Offenbach, Germany.

*CBW and GA contributed equally to this article.
Commercial relationships: None.
Corresponding author: Guillermo Aguilar.
Email: guillermo@bccn-berlin.de.
Address: Modeling of Cognitive Processes, Technische Universität Berlin and Bernstein Center for Computational Neuroscience, Berlin, Germany.

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed). (pp. 339–351). Cambridge, MA: MIT Press.

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1):37, 1–18, doi: 10.1167/17.1.37. [PubMed] [Article]

Anderson, B. L. (1999). Stereoscopic surface perception. *Neuron*, *24*(4), 919–928.

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978–R983.

Arend, L. E., & Goldstein, R. (1987). Simultaneous constancy, lightness, and brightness. *Journal of the Optical Society of America A*, *4*(12), 2281–2285.

Baird, J. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Barrow, H. G., & Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman (Eds.), *Computer vision systems* (pp. 3–26). New York: Academic Press.

Brainard, D., Brunt, W., & Speigle, J. (1997). Color constancy in the nearly natural image. I. Asymmetric matches. *Journal of the Optical Society of America A*, *14*, 2091–2110.

Chubb, C., Landy, M. S., & Econopouly, J. (2004). A visual mechanism tuned to black. *Vision Research*, *44*(27), 3223–3232.

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3):19, 1–14, doi:10.1167/12.3.19. [PubMed] [Article]

D'Zmura, M., & Iverson, G. (1993). Color constancy. II. Results for two-stage linear recovery of spectral descriptions for lights and surfaces. *Journal of the Optical Society of America A*, *10*(10), 2166–2180.

Ekroll, V., & Faul, F. (2013). Transparency perception: The key to understanding simultaneous color contrast. *Journal of the Optical Society of America A*, *30*, 342–352.

Fechner, G. (1860). *Elemente der psychophysik* [Translation: *Elements of psychophysics*]. Leipzig, Germany: Breitkopf und Hartel.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, *94*, 62–75.

Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological Science*, *22*(6), 812–820.

Foster, D. (2003). Does colour constancy exist? *Trends in Cognitive Sciences*, *7*, 439–443.

Foster, D. (2011). Color constancy. *Vision Research*, *51*(7), 674–700.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Review of Psychology*, *39*, 169–200.

Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., ... Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, *106*(4), 795–834.

Hillis, J., & Brainard, D. (2007). Distinct mechanisms mediate visual detection and identification. *Current Biology*, *17*, 1714–1719.

Kingdom, F., & Prins, N. (2010). *Psychophysics: A practical introduction*. London: Academic Press.

Knoblauch, K., & Maloney, L. T. (2008). MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software*, *25*, 1–26.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York: Springer.

Koenderink, J. (2013). Methodological background: Experimental phenomenology. In J. Wagemans (Ed.), *Handbook of perceptual organization* (pp. 41–54). Oxford, UK: Oxford University Press.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*. volume I: Additive and polynomial representations. Mineola, NY: Dover Publications.

Logvinenko, A. D., & Maloney, L. T. (2006). The proximity structure of achromatic surface colors and the impossibility of asymmetric lightness matching. *Perception & Psychophysics*, *68*(1), 76–83.

Logvinenko, A. D., Petrini, K., & Maloney, L. T. (2008). A scaling analysis of the snake lightness illusion. *Perception & Psychophysics*, *70*(5), 828–840.

Lu, Z.-L., & Sperling, G. (2012). Black-white asymmetry in visual perception. *Journal of Vision*, *12*(10):8, 1–21, doi:10.1167/12.10.8. [PubMed] [Article]

Maertens, M., & Shapley, R. (2013). Linking appearance to neural activity through the study of the perception of lightness in naturalistic contexts. *Visual Neuroscience*, *30*, 289–298.

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8):5, 573–585, doi:10.1167/3.8.5. [PubMed] [Article]

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913.

Munsell, A. E. O., Sloan, L. L., & Godlove, I. H. (1933). Neutral value scales. I. Munsell neutral value scale. *Journal of the Optical Society of America*, *23*(11), 394–411.

Obein, G., Knoblauch, K., & Viénot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision*, *4*(9):4, 711–720, doi:10.1167/4.9.4. [PubMed] [Article]

Pauli, H. (1976). Proposed extension of the CIE recommendation on uniform color spaces, color difference equations, and metric color terms. *Journal of the Optical Society of America*, *66*(8), 866–867.

Radonjić, A., & Brainard, D. H. (2016). The nature of instructional effects in color constancy. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 847–865.

Radonjić, A., Cottaris, N., & Brainard, D. (2015a). Color constancy in a naturalistic, goal-directed task. *Journal of Vision*, *15*(13):3, 1–21, doi:10.1167/15.13.3. [PubMed] [Article]

Radonjić, A., Cottaris, N., & Brainard, D. (2015b). Color constancy supports cross-illumination color selection. *Journal of Vision*, *15*(6):13, 1–19, doi:10.1167/15.6.13. [PubMed] [Article]

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R*. New York: Springer.

Singh, M. (2004). Lightness constancy through transparency: Internal consistency in layered surface representations. *Vision Research*, *44*, 1827–1842.

Singh, M., & Anderson, B. L. (2002). Toward a perceptual theory of transparency. *Psychological Review*, *109*(3), 492–519.

Umbach, N. (2013). Dimensionality of the perceptual space of achromatic surface colors. (Doctoral dissertation, Eberhard-Karls-Universität Tubingen, presented March, 2014; published May, 2014). *Verlag Dr. Hut*.

Whittle, P. (1994). The psychophysics of contrast brightness. In A. L. Gilchrist (Ed.), *Lightness, brightness, and transparency* (pp. 35–110). New York: Psychology Press.

Wiebel, C. B., Singh, M., & Maertens, M. (2016). Testing the role of Michelson contrast for the perception of surface lightness. *Journal of Vision*, *16*(11):17, 1–19, doi:10.1167/16.11.17. [PubMed] [Article]

Zeiner, K., & Maertens, M. (2014). Linking luminance and lightness by global contrast normalization. *Journal of Vision*, *14*(7):3, 1–15, doi:10.1167/14.7.3. [PubMed] [Article]