Behavior Research Methods, Instruments, & Computers
2003, 35 (1), 70-81

Measuring surface achromatic color: Toward a common measure for increments and decrements

GÁBOR JANDÓ
University of Pécs, Pécs, Hungary

and

TIZIANO AGOSTINI, ALESSANDRA GALMONTE, and NICOLA BRUNO
University of Trieste, Trieste, Italy

Surface color is traditionally measured by matching methods. However, in some conditions, the color of certain surfaces cannot be measured: The surface simply looks brighter or darker than all the patches on a matching scale. We studied the reliability, validity, and range of application of three different types of simulated Munsell scales (white-, black-, and split-surrounded) as methods for measuring surface colors in simple disk–ring displays. All the scales were equally reliable for matching both increments and decrements, but about 20% of the increments were unmatchable on the white-surrounded scale, about 13% of the decrements were unmatchable on the black-surrounded scale, and about 9% of the increments were unmatchable on the split-surrounded scale. However, matches on all the scales were linearly related. Therefore, it is possible to convert them to common units, using regression parameters. These units provide an extended metric for measuring all increments and decrements in the stimulus space, effectively removing ceiling and floor effects, and providing measures even for surfaces that were perceived as out of range on some of the scales.

Proper quantification of appearance is fundamental in any study on visual perception. In this paper, we report findings that bear on the issue of the quantification of achromatic color. Achromatic color is traditionally measured using matching paradigms, which require using an adjustable patch or a set of patches (a gray scale) that an observer can inspect to choose a match. However, it is known that this method will not provide a match for a surface in general. In some conditions, some surfaces will look either too dark or too light for any of the matching patches on a given scale. This problem involves several issues that have been touched on previously in the literature, including the difficulty of matching increments to decrements, the possibility of a qualitative change from matches of perceived reflectance (*lightness*) to matches of perceived luminance (*brightness*), and the putative appearance of *superwhites*, surfaces that look both self-emitting and opaque. In this paper, we propose a method for computing a unified metric for achromatic surface color, encompassing matches with different scales and providing a new view of the above issues. In other research (Bruno, Jandó, Galmonte, & Agostini, 1998), we have applied this method to a study of surfaces surrounded by linear gradients.

MEASURING SURFACE COLOR

It is known, at least since the seminal work of Kardos (1934) and Katz (1935), that surface color is a function of all the surfaces in a given scene. Kardos wrote this function as

$$C(\mathrm{T}) = f(s_1, s_2, s_3, \ldots s_i), \qquad (1)$$

where $C(\mathrm{T})$ is the color of a target surface, s_i is the luminance of the ith surface, and f means simply that a certain stimulus surface corresponds to a position or a set of positions in a multidimensional space of the given parameters. This is only a formal definition, not a quantitative description. It amounts to assuming that observers will perceive similar things while looking at the same image and that the perception of a color is fairly stable in the time domain. None of the above statements is entirely true, given individual differences and the potential effects of age, drugs, training, and perceptual set.

While most of these effects are commonly believed to be weak or transitory, several investigators believe that, at least,

This work was supported by OTKA T020428, OTKA T023657, a Berde fellowship, and a scholarship of the "Consorzio per lo Sviluppo Internazionale dell'Università di Trieste," all to the first author. Additional support was provided by MURST Grants MM11197484 and MM534119 to the second and the last authors, respectively. We are grateful to Paolo Bernardis, Igor Castellarin, Fauzia Mosca, and Caterina Ripamonti for the time spent performing matches in the two experiments. Finally, we thank Ira H. Bernstein, Joseph Cataliotti, and Jonathan Vaughan for helping us, with their constructive comments, to improve the quality of our paper. Correspondence concerning this article should be addressed to T. Agostini, Dipartimento di Psicologia, Università di Trieste, via S. Anastasio, 12, 34134 Trieste, Italy (e-mail: agostini@univ.trieste.it).

the set of an observer is important. In this view, whenever the surfaces to be matched can be perceived as being under different illuminations, two different perceptual dimensions can be matched: perceived reflectance, or lightness, which is approximately invariant with changes in illumination, and perceived luminance, or brightness, which varies as a function of a change in illumination (for various treatments of these issues, see the chapters in Gilchrist, 1994). Since Wallach (1948), several investigators have proposed that matching a surface on a decremental scale (i.e., a scale of Munsell papers surrounded by white) tends to elicit lightness matches, whereas matching on an incremental scale (i.e., a scale of Munsell papers surrounded by black) tends to elicit brightness matches (Gilchrist, 1988; Heinemann, 1989; Jacobsen & Gilchrist, 1988b). However, to our knowledge, this claim has never been tested systematically. What seems to be true, as has been verified by several studies, is that certain surfaces in certain conditions cannot be matched at all. For instance, it is known that, in simple displays, increments cannot be matched to decrements (Whittle & Challands, 1969). Unmatchable surfaces have sometimes been reported when a simple disk–ring arrangement has been compared with a more complex stimulus (Bruno, 1992; Bruno, Bernardis, & Schirillo, 1997). Most important, a number of investigators have noted that within the same stimulus space, some surfaces can appear definitely too bright, almost self-emitting, for matching on a decremental scale to be possible (Bonato & Gilchrist, 1994; Heinemann, 1955; MacLeod, 1947). This peculiar mode of appearance has been called *fluorescence* (Evans, 1959, 1974) or, more recently, *superwhite* (Gilchrist et al., 1999). Attempts at constructing a comprehensive psychophysics of achromatic color matches have also identified parameters that yield a sort of superblack region—that is, luminance ratios that yield appearances that cannot be predicted on the basis of a ratio principle (see, for instance, Whittle's chapter in Gilchrist, 1994).

Whatever the interpretation of these observations, a measuring problem remains. To assess surface color in a generic image, one need a unified metric for surfaces surrounded by any kind of background (see Figure 1). Given the above difficulties, it is not at all clear what kind of matching scale is appropriate for computing this metric. In a number of studies, decremental scales have been used for lightness matches and incremental scales for brightness matches. However, this forces an investigator to restrict the range of studied displays, for some surfaces will

$$M(\mathrm{T}) = f(s_1, s_2, s_3, s_4 \ldots s_n)$$

Figure 1. In a generic image, a target surface T can be surrounded completely or in part by surfaces that can be either darker or lighter. The simple fact that T can be a luminance increment, a decrement, or both poses a fundamental problem for measuring its surface color, using a matching paradigm. If a scale is needed to select a matching patch, what is the proper surround for this scale?

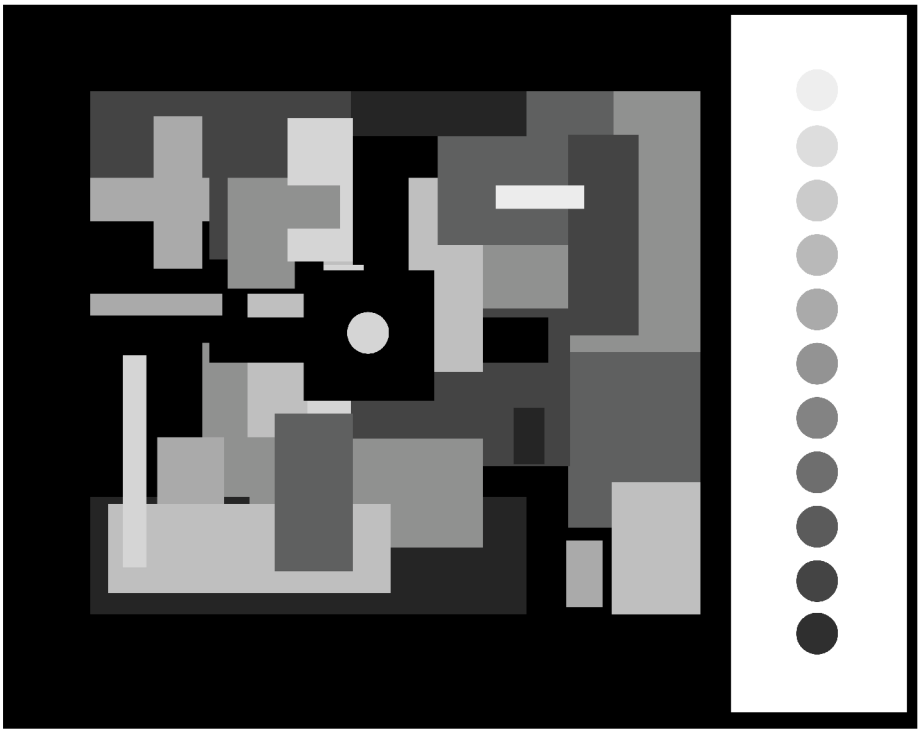

Figure 2. Illustration of the out-of-range problem in matches of achromatic color. Note that the central surface looks too bright for any of the patches in the matching scale.

be impossible to match. Other investigators have used scales presented on black-and-white checkerboards (for instance, Schirillo & Shevell, 1996). However, even with checkerboard-surrounded scales, unmatchable surfaces can occur. In addition, it is not well understood how these matches relate to those on simple incremental or decremental scales. Finally, in several published works, Gilchrist (Jacobsen & Gilchrist, 1988a, 1988b) collected matches on a hybrid scale consisting of a set of decremental patches on a white surround plus an adjustable self-emitting region, also on white. The self-emitting region, constructed by adding a small chamber containing an adjustable light sending diffuse illumination through a translucent patch, effectively extended the lightness scale by providing a means for matching surfaces that were too bright for any of the decremental patches. Although interesting, Gilchrist's approach relies on the untested assumption that matches on the incremental, emitting patch will form a perceptual continuum with the decremental patches. This assumption is challenged by a number of other papers, some by Gilchrist himself (Jacobsen & Gilchrist, 1988a), suggesting that lightness and brightness are separate continua. In addition, it is not clear how the hybrid scale could be applied on displays that appear under a homogeneous illumination so that perceived reflectance and perceived luminance are completely correlated. In the present paper, we demonstrate that unmatchable surfaces do occur in these conditions.

AN EXTENDED SCALE FOR INCREMENTS AND DECREMENTS

Whenever one attempts to measure physical, chemical, biological, or psychological quantities, it is crucial to obtain measurements that can be compared across methods of measurement. Scales and units in general are based on reasonable but always arbitrary conventions. However, the most important common requirement for any quantitative method is that the measurements be reproducible and that the method be set up in such a way that the discrepancy between two repeated measurements would be the least possible. Theoretically, a measurement can be considered as an evaluation of a function, where the independent variable is the underlying variable to be measured, the function captures processes that can cause an observable change in a system (i.e., changing an index, moving a pin, etc.), and the dependent variable is the result of the measurement (i.e., the degree of shift). If the function is known, the underlying physical variable can be calculated from the inverse of the function. In certain types of measurement, one can even neglect the underlying variable and simply define what is being measured by the very method of mea-

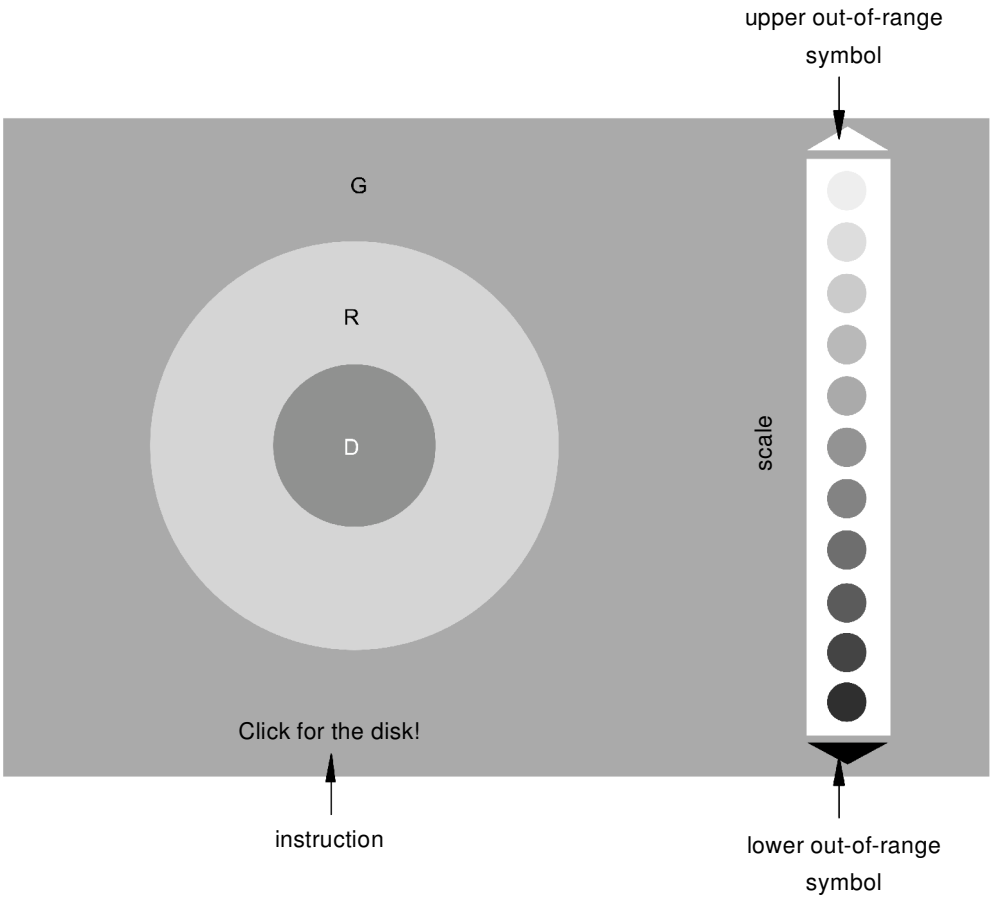

Figure 3. Experimental display. A disk (D) surrounded by a ring (R) was placed at the center of the monitor. On the right side of the monitor, a vertically oriented Munsell scale was simulated. The general background (G) was set to a constant value.

surement. The measurement of temperature, for example, is treated in this way, outside of molecular physics. The volume of a solid mass, liquid, or gas depends on its temperature. Thus, a thermometer can exploit this fact even if temperature has nothing to do with volume. The essence of temperature can be found somewhere at the level of molecular motion, but we do not want to calculate molecular motion when we measure temperature. Instead, temperature is defined as a certain state of a certain material (i.e., mercury) in certain conditions (i.e., a tube).

Any measurement can differ from any other measurement either because of random error or because of systematic biases. The straightforward approach to reducing random error is to average several measurements. Systematic biases, on the other hand, can be due to confounding variables or to the very measuring method being used. For instance, temperature readings on a Celsius thermometer will differ from those on a Fahrenheit thermometer because the measuring systems are different. Nonetheless, the two measures can still be used in conjunction if there is a way of converting one to the other or converting both to a third, common measure (i.e., the Kelvin scale). Once the measurements are all in similar units, it becomes possible to compare them. Suppose that two different measures are the following functions of the underlying mechanism:

$$M = f(u) \tag{2}$$

and

$$N = g(u), \tag{3}$$

where M and N are the two different measures, and f and g are two different functions of the same underlying mechanism (u). If we would like to know the function

$$N = h(M), \tag{4}$$

which allows us to convert one measure into the other, it is easy to prove that

$$N = h(M) = g[f^{-1}(M)]. \tag{5}$$

Thus, any pair of measuring scales that are in unequivocal congruence with the same underlying variable (i.e., that

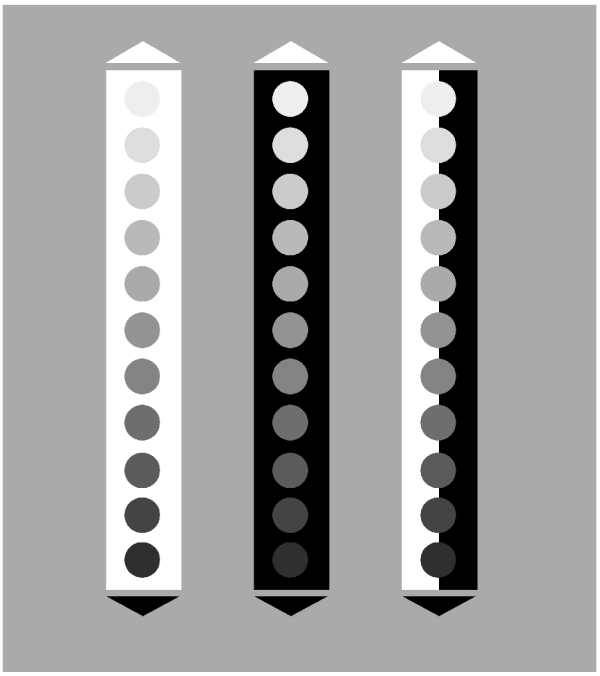

Figure 4. Schematics of the three scales (corresponding patches have the same reflectance). Scales actually used in the experiments had 32 different patches.

have inverse functions) can be converted to each other, and the resulting combined scale can be used for quantification. If we do not know the *f* and *g* functions but can make measurements with both methods, the following procedure can be used to determine *h*. First, take measurements in different conditions (as many conditions as possible) with both methods. (The conditions should be within the range that we would like to study.) Second, plot the data measured by Method 1 against the value measured by Method 2. Third, establish the function between the two scales by fitting the most suitable function to the data. Finally, use the function to convert the data to comparable units.

We speculated that a similar approach could be used for measuring surface achromatic color. As was stated above, none of the existing matching methods is fully satisfactory. For instance, a high-contrast incremental surface cannot typically be matched on a decremental scale, such as a Munsell scale on a white surround, which is typically used to obtain lightness matches (see Figure 2). A high-contrast decrement will tend to give similar problems when matched on an incremental scale. If, however, each of these matching scales can be considered as measuring methods that are valid within a certain range and if well-behaved relationships can be established between measurements on each scale, it may be possible to convert them all to a common, extended scale. This scale would then provide a unified metric for any surface in the stimulus space.

GENERAL METHOD

To investigate matches of surface achromatic color as a function of the type of matching scale, two experiments were performed. In the first experiment, the test–retest reliability of the matches within each scale was established by presenting the same displays in two successive sessions. In the second experiment, matches on the three different scales were compared within each image to construct an extended achromatic–color scale for increments and decrements. All the displays were simulated on a computer monitor and presented on the same general background in order to evoke an impression of a single, homogenous illumination on the whole simulated pattern.

Observers

Eight observers participated in both experiments. Three of them (T.A., A.G., and N.B.) were experienced in the psychophysics of achromatic surface color, whereas the remaining 5 were completely naive.

Equipment

All the stimuli were generated using a Silicon Graphics Indigo workstation and were displayed on a carefully calibrated Silicon Graphics monitor. This monitor has a resolution of 1,280 × 1,024 pixels and 256 simultaneously displayable gray levels covering a range of approximately 2 log units of luminance. Monitor calibration was performed in two steps. First, photometer readings were obtained for the darkest and the brightest grays that could be produced on the monitor, and the contrast and brightness switches were adjusted to achieve a range of about 2 log units. Next, luminances at different monitor locations and at different gray values were measured for each software-specifiable gray level (that is, 0 to 255), and the resulting 256 luminance–gray-level pairs were stored in a look-up table. Finally, luminances were converted to relative luminance val-

Table 1
Reliability Coefficient for Each of the Scales

	Observers																	
	A.G.		F.M.		G.J.		I.C.		C.R.		N.B.		P.B.		T.A.		Average	
Scale Type	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*	*n*	*r*
								Increments										
Black	20	.98	17	.89	18	.97	19	.98	23	.94	28	.97	22	.94	23	.95	21	.95
White	20	.92	17	.89	18	.99	19	.95	23	.93	28	.97	22	.97	23	.85	21	.93
Split	20	.98	17	.88	18	.98	19	.97	23	.93	28	.99	22	.88	23	.95	21	.95
								Decrements										
Black	20	.82	23	.94	22	.93	21	.96	17	.95	32	.91	18	.96	17	.94	21	.93
White	20	.97	23	.94	22	.89	21	.90	17	.96	32	.95	18	.88	17	.98	21	.93
Split	20	.96	23	.96	22	.96	21	.95	17	.95	32	.83	18	.92	17	.91	21	.93

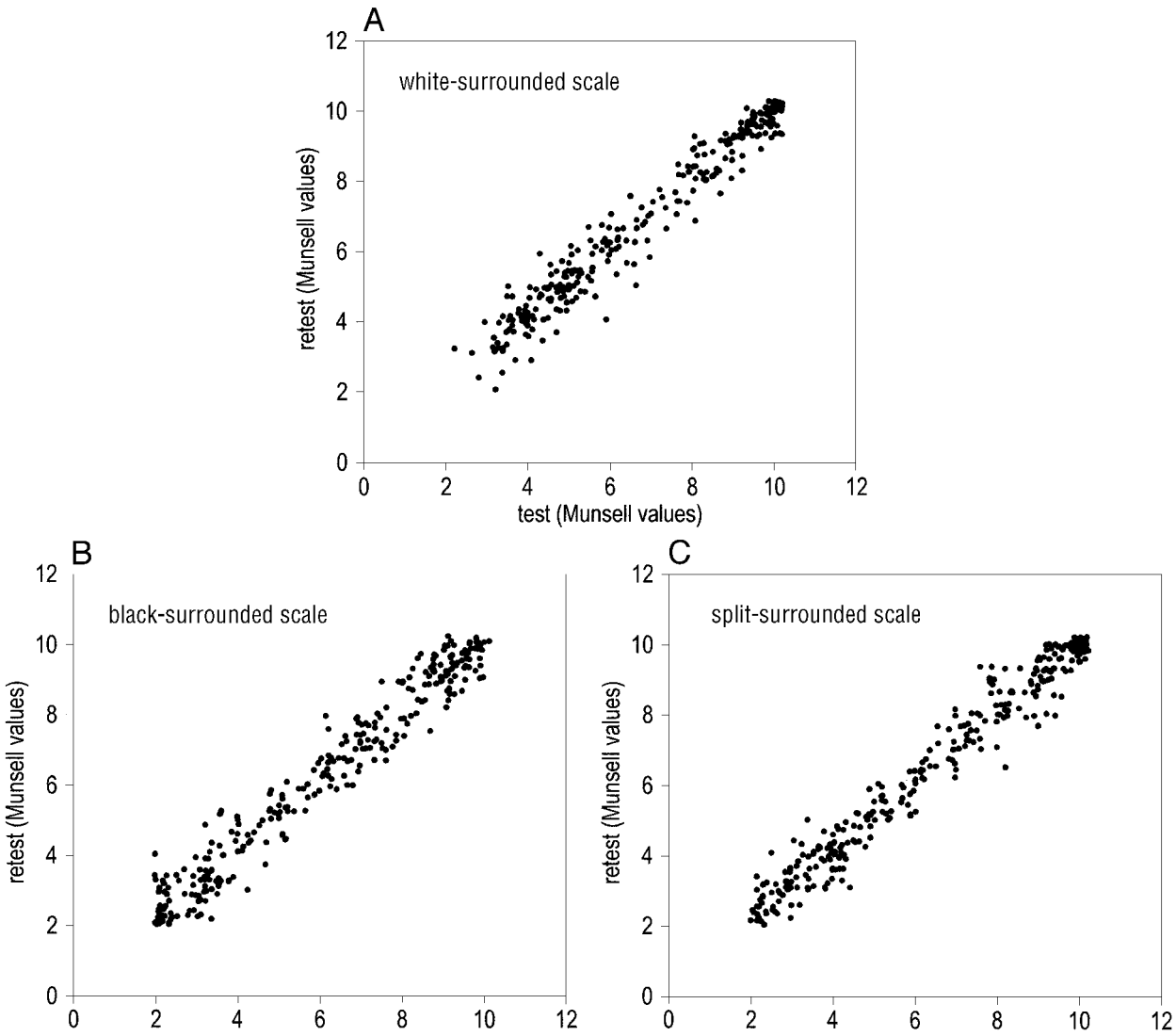

Figure 5. Test–retest correlations in the three scales. All 8 observers are plotted simultaneously. Each data point represents a pair of matches on the same image by the same observer.

ues (RL) ranging from 0 to 1 (gray 0 = 0.000 RL = 0.2 cd m^2; gray 255 = 1.000 RL = 60 cd m^2). An RL value of 0 was defined as a simulated Munsell value of 0/, whereas an RL of 1 was defined as a simulated value of 10/. Intermediate values were assigned a simulated Munsell value according to the conversion table of Judd (1966).

Displays

Experimental displays were presented on the computer monitor in a completely dark room (see Figure 3). They consisted of a central disk surrounded by a ring, both of variable luminance, set against a general background of constant luminance equal to 18.2 cd m^2. To ensure homogeneous sampling of log-ratio space, luminances were assigned to the disk and the ring according to the following procedure. First, a value was selected at random, with replacement in the range between log (0.1/0.85) and log (0.85/0.1). Second, a log-relative luminance was chosen at random in the range between log (0.1) and log (0.85). Third, the second log-relative luminance was computed from the ratio selected in Step 1. If the second relative luminance was below 0.1 or above 0.85, Step 2 was repeated until a suitable pair was chosen. The diameter of the central disk subtended 3.75° of visual angle, the diameter of the surrounding ring subtended 12°, and the horizontal extent of the general background was equal to

Table 2
Frequencies (*F*) of Out-of-Range Judgments in the Three Scales With Increments and Decrements and Percentages Relative to 340 Matched Surfaces Within Each Scale, Rounded to the Second Decimal Digit

	Scale					
	Black		White		Split	
Luminance Contrast	*F*	%	*F*	%	*F*	%
Increments	6	2	70	20	30	9
Decrements	44	13	0	0	6	2

Table 3
Frequency (*F*) of Upper and Lower Out-of-Range Judgments With Increments and Decrements and Percentages Relative to 1,020 Matched Surfaces Within All Scales, Rounded to the Second Decimal Digit

	Increments		Decrements	
Judgments	*F*	%	*F*	%
	Upper			
Consistent	71	7	0	0
Inconsistent	35	3	0	0
	Lower			
Consistent	0	0	37	4
Inconsistent	0	0	13	1

26.25°, at a viewing distance of approximately 76 cm. Thus, the whole display filled a larger portion of the observer's visual field, minimizing unwanted interactions with the uncontrolled, dark surround around the monitor (Agostini & Bruno, 1996; Agostini & Galmonte, 1999). To the right of the disk–ring arrangement and next to the right edge of the monitor, a simulated Munsell scale (32 patches from simulated Munsell 2/ to Munsell 9.75/, in steps of 0.25/) was presented, as is illustrated in Figure 3. Each randomly selected display was presented once with a scale on a white surround (simulated Munsell value 10/), once with a scale on a black surround (0.3/), and once with a scale on a split surround (left half 10/, right half 0.3/; see Figure 4)[1]. In addition, each scale included a black triangle at the bottom and a white triangle on the top. These were used to signal that the displayed achromatic color was out of the range of the given scale.

General Procedure

Matches were performed in sessions consisting of 10 randomly selected disk–ring arrangements. Each was paired with each of the scales. The resulting 30 displays were shown in completely randomized order. The observers were instructed to match the achromatic colors of the disk and of the ring by clicking the appropriate patches on the available scales, yielding a total of 60 matches per session. The 2 matches (disk and ring) for each image were also performed in random order by prompting the subject for a disk or a ring match according to a computer-simulated coin toss in the experimental program. If no match was possible, observers were asked to click on the lower (color too dark for any of the patches) or upper (color too bright) out-of-range symbols. Each session lasted about 5 min. A break was mandatory between sessions to avoid adaptation effects.

Instructions

The exact instructions read to the observers were, "Find the gray patch on the scale that you think would melt into the color of the area in question if you could detach it from the scale and move it over the area to be matched. If the color looks too bright for any of the patches, click on the upper triangle. If the color looks too dark, click on the lower triangle."

EXPERIMENT 1
Test-Retest Reliability of Achromatic Color Matches

Method

Procedure and Analysis. All the observers performed a total of four sessions [two tests and two retests; in order: test(1), test(2), few days pause, retest(1), retest(2)], except for Observer N.B., who performed six sessions. Each retest session was completed at least 1 day after the test and no later than 3 days after. Each pair of sessions was administered in the same random order. Matches in the first session were compared with those in the second session by computing separate test–retest reliabilities for the black-, white-, and split-surrounded matching scales (henceforth, the *black*, *white* and *split* scales) and for incremental and decremental disk–ring arrangements.

Results and Discussion

Table 1 summarizes test–retest correlations for each observer, matching increments and decrements with each of the three scales. Increments and decrements were defined in relation to the given surrounding luminance. Thus, the disk was an increment or decrement relative to the ring, whereas the ring was an increment or decrement relative to the general background, independent of its relation to the disk. Computation of the individual correlations included the out-of-range judgments, coded as a match equal to 0 for the lower out-of-range judgments and a match of 10 for the upper out-of-range judgments.

Figure 5 plots matches in the first session against matches in the second session for each of the three scales and for all observers, excluding out-of-range judgments. Individual reliability coefficients were all highly satisfactory, ranging from $r(19) = .82$ to $r(27) = .99$, without any systematic difference between scales or between matches to increments and decrements.

Tables 2, 3, and 4 summarize frequencies of out-of-range judgments as a function of type of scale, increments and decrements, and consistency. Several things are noteworthy. As is shown in Table 2, essentially all of out-of-range judgments on the white and the split scales occurred with increments, whereas most out-of-range judgments on the black scale occurred with decrements. As is shown in Table 3, upper out-of-range judgments occurred twice as often as lower out-of-range judgments. Most likely, this asymmetry was due to the fact that the minimum relative luminance value in the stimuli was 0.1 (approximately, a Munsell value of 3.6/), whereas the maximum was 0.85 (approximately 9.5/). Note also that all upper out-of-range judgments occurred with increments, whereas all lower out-of-range judgments occurred with decrements. Table 3 also demonstrates that the observers were fairly consistent when providing out-of-range judgments. In Table 4, frequencies of consistent and inconsistent out-of-range judgments are further subdivided by matching scale. Interest-

Table 4
Frequency (*F*) of Upper and Lower Out-of-Range Judgments in the Three Scales and Percentages Relative to 340 Matched Surfaces Within Each Scale, Rounded to the Second Decimal Digit

	Scale					
	Black		White		Split	
Judgment	*F*	%	*F*	%	*F*	%
	Upper					
Consistent	5	1	57	17	13	4
Inconsistent	1	0	13	4	17	1
	Lower					
Consistent	36	11	0	0	1	0
Inconsistent	8	2	0	0	5	1

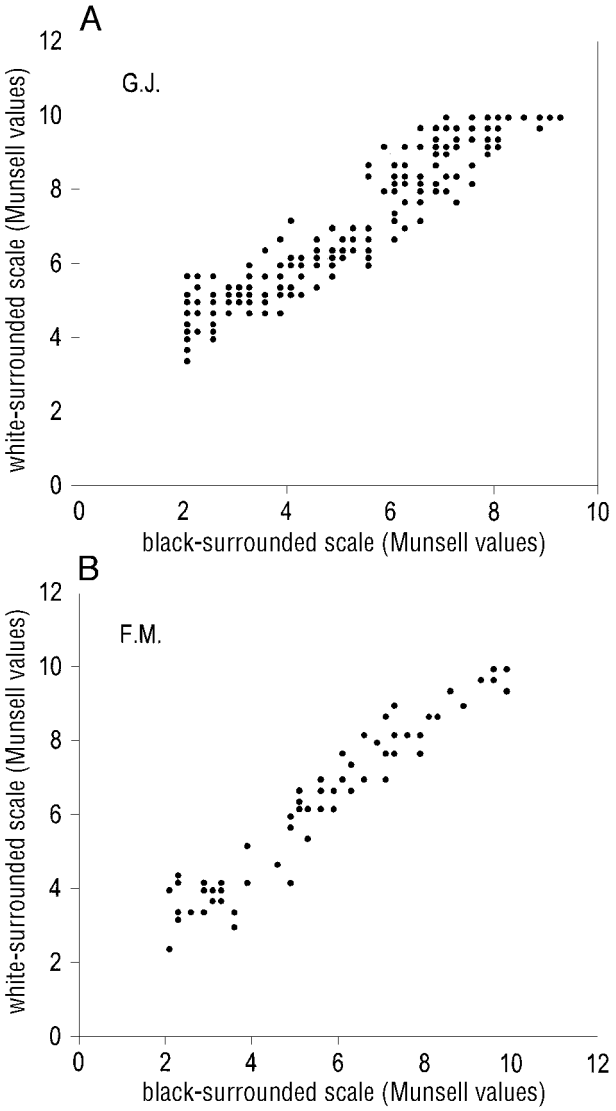

Figure 6. Linear correlation between matches on the white- and black-surrounded scales in 2 representative observers. The corresponding regression parameters can be found in Table 2.

ingly, the major contributor to the inconsistent judgments was the split scale, which in fact yielded more inconsistent cases than consistent ones. A test of independence confirmed that this pattern was statistically significant [$\chi^2(2) = 25.03, p < .001$].

Overall, these results demystify a common belief about the matching method for measuring achromatic color. On the basis of our observed reliabilities, there are little grounds for the claims that the white scale is most reliable for decrements, whereas the black scale is most reliable for increments, or that an incremental scale is generally more variable than scales including a white anchor. The main difference between the scales appears to be in the location of the out-of-range surfaces in the space of the selected displays.

EXPERIMENT 2
Combining Matches on Different Scales

Method

Procedure and Analysis. All the observers completed 4 sessions on different, randomly selected displays, except for observer G.J., who completed 22 sessions. Matches on the same images were compared across scales in order to establish whether surfaces that could not be matched on one scale could be matched on another and whether matches from different scales could be combined into an extended achromatic color scale. More precisely, for each observer, we plotted matches on the white scale against matches on the black scale and matches on the white scale against matches on the split scale. Next, we determined by regression analysis whether the matches on different scales were linearly related. Given that they clearly did, we computed individual linear regression parameters to convert

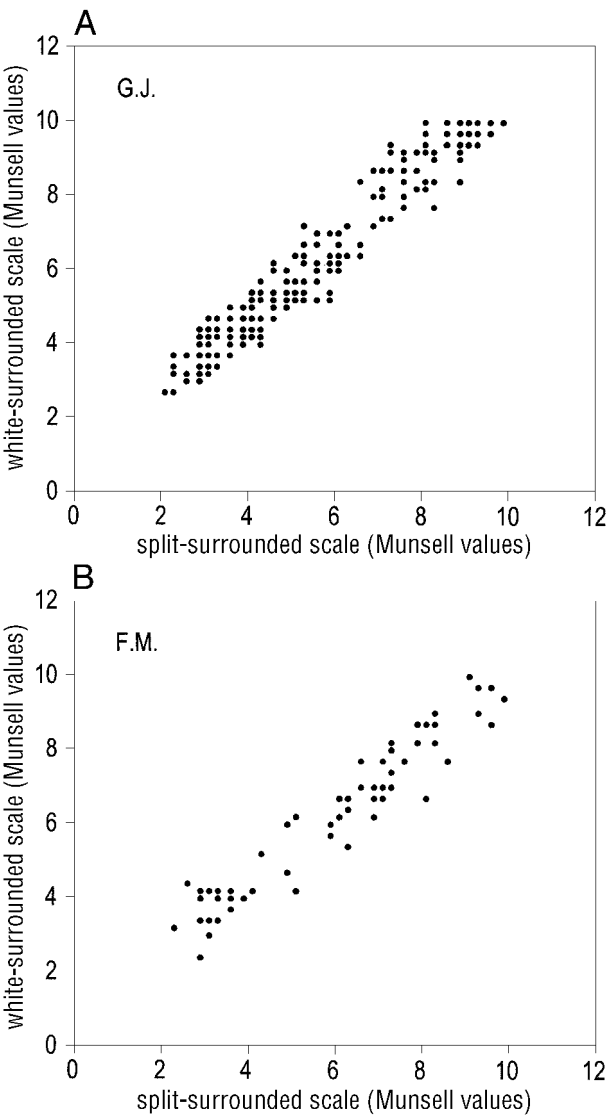

Figure 7. Linear correlation between matches on the white- and split-surrounded scales for 2 representative observers. The corresponding regression parameters can be found in Tables 5 and 6.

Table 5
Individual Linear Regressions of the White-Surrounded Scale on the Black-Surrounded Scale

Subject	Slope ± *SE*	Constant ± *SE*	r^2
A.G.	1.079 ± 0.098	2.745 ± 0.404	.770
F.M.	0.879 ± 0.033	1.312 ± 0.194	.912
G.J.	0.903 ± 0.016	2.224 ± 0.084	.924
I.C.	0.780 ± 0.032	2.324 ± 0.199	.896
C.R.	0.953 ± 0.035	2.139 ± 0.199	.922
N.B.	0.894 ± 0.048	3.400 ± 0.218	.843
P.B.	0.796 ± 0.023	2.356 ± 0.153	.937
T.A.	0.824 ± 0.068	3.874 ± 0.368	.781
Average	0.888	2.55	

matches on the black and the split scales to the same perceptual units as those used for matches on the white scale. Linear relationship between scales was also confirmed by an analysis of residuals (i.e., subtracting data from the regression equations). The residuals from upper values were compared with the residuals from lower values with a one-way analysis of variance and a *t* test. Finally, we compared the resulting extended-unit scale with the three individual scales within the context of a simple model predicting the absolute achromatic color of a surface as a function of two parameters, the ratio of the luminance of the ring to the luminance of the disk, and the ratio of the luminance of the general background to the luminance of the disk.

Results

Figure 6 plots matches of the same surface on the white scale against corresponding matches on the black scale, for 2 representative observers. Figure 7 presents the same plot for matches on the white scale against corresponding matches on the split scale. Both plots do not include out-of-range judgments. Two things are noteworthy. First, there is a clear linear relationship between the matches on different scales. More precisely, matches of a given surface on the white scale can be predicted from the matches on the black scale by adding approximately 2 Munsell units. Conversely, matches on the white scale can be predicted from the matches on the split scale by adding approximately 1 Munsell unit. More exact conversions can be performed for all individual observers or for an ideal average observer by using the regression parameters provided in Tables 5 and 6. The analysis of residuals confirmed, that the relationship between scales is truly linear, since data points were evenly distributed around the regression equation with no bias. Second, the bivariate spread of the data is comparable to that of the test–retest plots, suggesting that the lack of a tighter relationship between matches on different scales is most likely due to random variability in the matches, not to any systematic difference between the scales.

Closer inspection of the individual regression parameters in Tables 5 and 6 reveals some individual differences. By comparing individual parameter estimates and their standard errors, it seems that all the individual slopes are essentially equivalent and consistent with ideal-observer slopes around 0.9, whereas individual *y*-intercepts show greater variability. In particular, the *y*-intercepts of 3 observers (A.G., N.B., and T.A.) seem to be significantly higher than those of the others. It is difficult, on the basis of the present data, to determine what might have caused this difference. Given that all the observers were between 25 and 35 years old, that they had no color vision abnormalities, to the best of their knowledge, and that no obvious sex difference was apparent, these factors are unlikely to have caused the difference. It may be noted that Observers A.G., N.B., and T.A. were also those that were experienced in color matching. As an informal test of the hypothesis that the difference was related to expertise, we asked two leading investigators to perform sessions in the experiment. Each of the latter observers had at least 20 years of experience in performing achromatic color matches. However, their individual estimates were essentially equivalent to those of our naive subjects. As a further check of long-term effects, we asked Observer G.J. to complete two more sessions of this study. At the beginning of the present study, G.J. had never performed color matches, but he later performed many additional matching sessions in a follow-up project. Next, we compared his parameter estimates from the first two sessions of the present data set with those of the additional sessions. The additional sessions were performed after about a month. Again, we did not find any shift in the computed parameters. These informal tests seem to rule out the possibility that the difference in individual intercepts was due to expertise.

Discussion

In general, it seems fair to say that matches from the three scales can be combined into an extended achromatic color scale, using regression equations. If we choose the white scale units as the baseline and the ideal-observer parameters as the best estimates of the relationships between the scales, the conversion of the three matches to extended units is provided by the equations

$$\text{Extended unit} = \text{white unit}, \quad (6)$$

$$\text{Extended unit} = 0.89 * \text{black unit} + 2.55, \quad (7)$$

and

$$\text{Extended unit} = 0.93 * \text{split unit} + 1.23. \quad (8)$$

After the conversion, it is possible to consider each match as a comparable assessment of surface achromatic color. For several surfaces, three such assessments are available. For some other surfaces, some assessments are lacking.

Table 6
Individual Linear Regressions of the White-Surrounded Scale on the Split-Surrounded Scale

Subjects	Slope ± *SE*	Constant ± *SE*	r^2
A.G.	1.075 ± 0.044	1.091 ± 0.224	.924
F.M.	0.889 ± 0.030	0.813 ± 0.187	.929
G.J.	0.982 ± 0.011	0.691 ± 0.066	.954
I.C.	0.917 ± 0.030	1.042 ± 0.194	.931
C.R.	0.959 ± 0.029	0.983 ± 0.195	.945
N.B.	0.857 ± 0.021	2.058 ± 0.118	.951
P.B.	0.851 ± 0.023	1.522 ± 0.160	.947
T.A.	0.946 ± 0.031	1.651 ± 0.171	.926
Average	0.934	1.23	

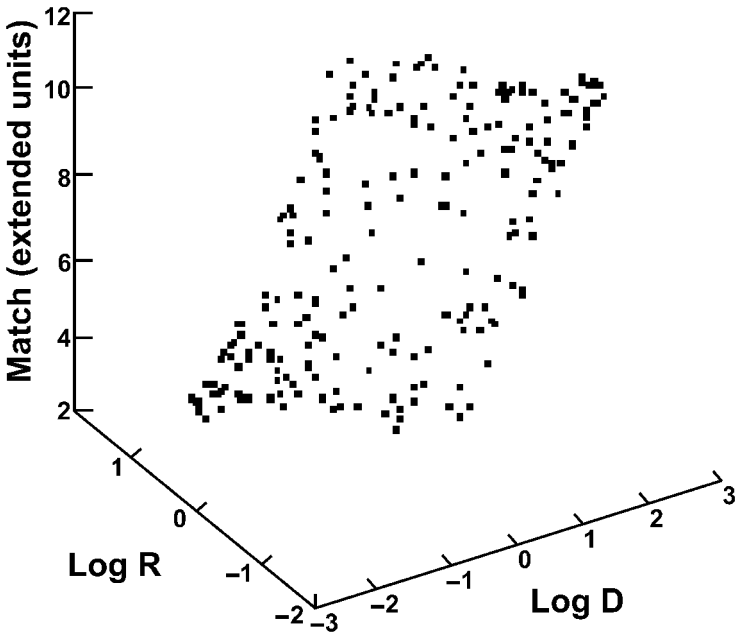

Figure 8. Observed matches for the disk (D) as a function of the log luminance of the ring (R) and the general background. Matches are expressed in extended units (see the text).

These are the cases in which some of the scales yielded *out-of-range* judgments. However, given that no surface ever yielded out-of-range judgments on all three scales, the resulting extended-unit scale encompasses the space of all possible disk–ring displays. To test the potential advantages of the extended-unit scale, we compared our matching data against the predictions of a simple model of achromatic surface color, for each of the scales separately, against the matches combined on the extended scale.

A large literature suggests that the achromatic color of a surface is a function of the luminance ratios between all the surfaces in a given scene (Jacobsen & Gilchrist, 1988b; Land & McCann, 1971; Wallach, 1948). In our simple displays, two such ratios are available: the ratio of the luminance of the ring to that of the disk and the ratio of the general background to that of the disk. Applying the ratio principle, we can express the relative achromatic color in terms of the following two equations:

$$m_r - m_d = k\,(\log r - \log d) \tag{9}$$

and

$$m_g - m_r = 1(\log g - \log r), \tag{10}$$

where r, d, and g are the luminances of the ring, the disk, and the general background, and m_r, m_d, and m_g are the corresponding matches. Combining Equations 9 and 10 yields an equation for the relative color of the disk to the general background:

$$m_g - m_d = k\,(\log r - \log d) + 1\,(\log g - \log r). \tag{11}$$

Finally, reordering Equation 11 yields an equation for the absolute color of a given surface. For instance, for the disk,

$$m_d = k\,(\log d - \log r) + 1\,(\log r - \log g) + m_g. \tag{12}$$

In terms of multiple regression, we can rewrite Equation 12 in a simpler way as follows:

$$m_d = a \log d + b \log r + c, \tag{13}$$

where the constants $a = k$, $b = 1-k$, and $c = -1 \log g + m_g$ are determined by regression analysis. On the basis of Equation 13, therefore, the achromatic color of the disk can be predicted from log d and log r if log g is constant. To visualize the model, three dimensions are needed, with the log d, log r plane representing the predictor variables and the expected arrangement of the data being a plane. The data from Experiment 2 are plotted in this way in Figure 8. Consistent with the implementation of the ratio principle of Equation 12, the data are indeed quite close to a plane.

To evaluate deviations of observed matches from the predictions of the model, we plotted observed matches against predicted matches. If the model predicts the achromatic color of the disk perfectly, all the points should lie on a line with a unitary slope. These plots are presented in Figure 9 for the observed matches in Munsell units, using the white, black, and split scales, and in Figure 10 for all the matches converted into extended units. The disadvantages of the individual scales are apparent. First, none of the individual scales can provide measurements for the whole range of displays, with the white scale faring worst

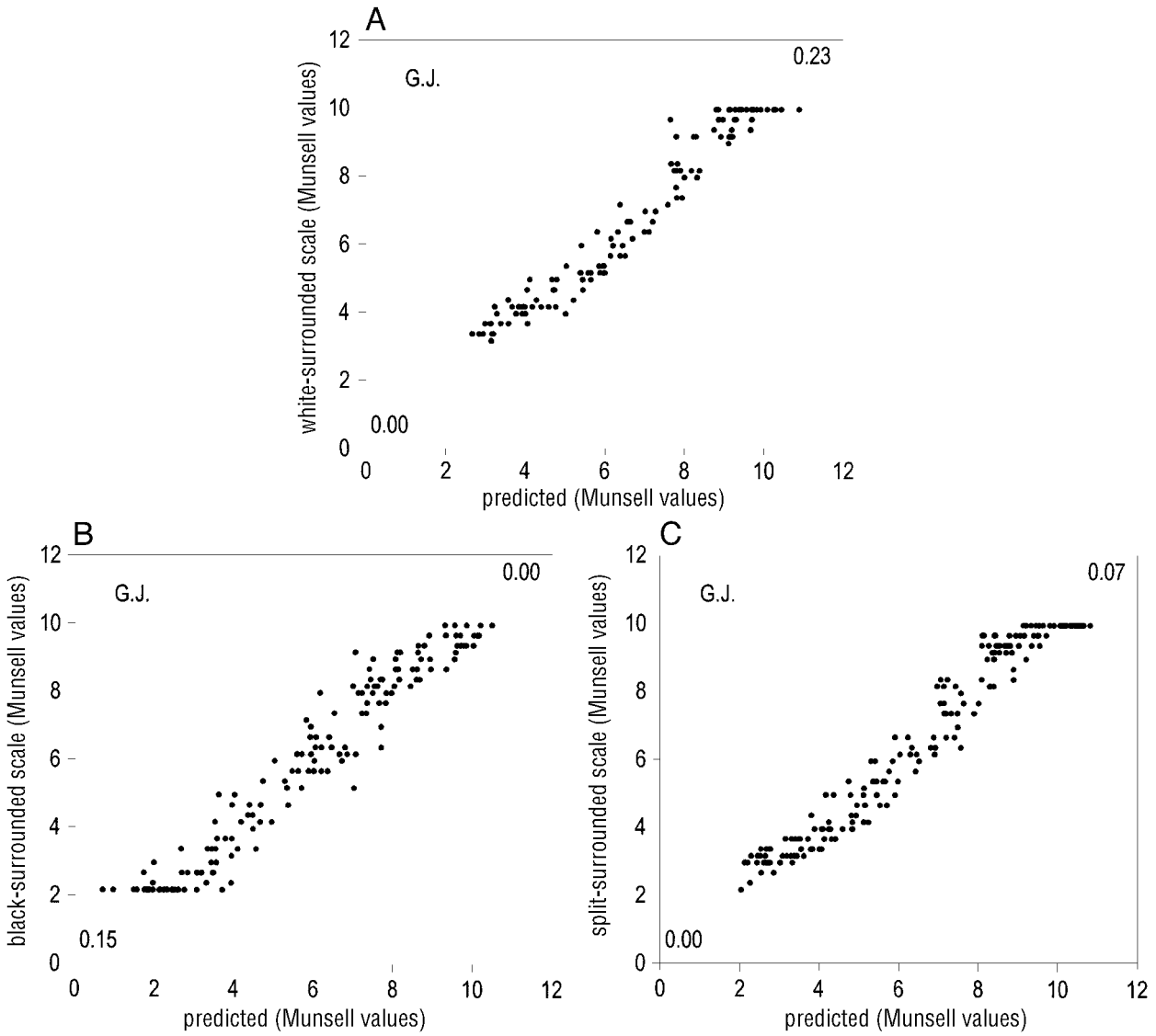

Figure 9. Observed matches (Munsell units) against predictions based on Equation 12 for each of the three scales. Lower left and upper right corners of each graph: proportion of lower and upper out-of-range judgments, out of 440. The data are from 1 representative observer.

(as much as 23% out-of-range judgments). Second, all of the individual scales exhibit clear ceiling (white and split scales) or floor (black scale) effects. Thus, these plots confirm that the white scale can provide measurements only in the lower-to-middle range of predicted achromatic colors, the black scale can provide measurements only in the upper-to-middle range, and the split scale appears to work best in the middle-to-lower range. These disadvantages are effectively solved by the conversion into extended units, which provides measurements and a common metric for all images.

CONCLUSIONS

Matching scales presented on different backgrounds provide equally reliable measurements of surface color within certain ranges. Outside of these ranges, individual scales appear to become inadequate, and often observers will report that a surface cannot be matched at all. However, also contrary to a commonly held view, matches obtained from different scales are comparable once they are converted to a common unit. That is, achromatic color scales presented on white, black, and split surrounds appear to behave very much like different scales for temperature. One can convert from one to the other once the conversion function is known. The present work exploited this fact to construct an ideal extended scale (obtained by converting, in a common metric, the values measured with individual scales) for achromatic color that allowed us to obtain reliable and valid measurements even for surfaces that are perceived as unmatchable on some of the individual scales. Using such an extended scale provides two impor-

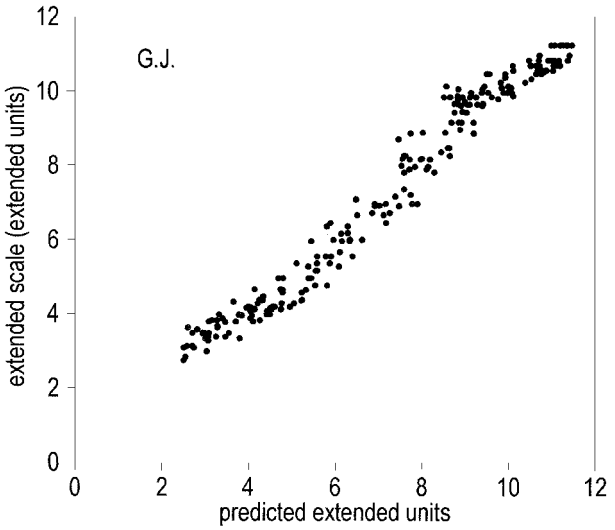

Figure 10. Observed matches (extended units) against predictions based on Equation 12. Given that no image yielded out-of-range judgments on all three scales, all the data can be plotted on the same graph, each point corresponding to a combined estimate of the match for a given surface. The data are from 1 representative observer.

tant advantages over using any of the scales individually. First, the extended scale provides a common metric for matches of any surface within a generic scene, even if the surface cannot be matched on one of the individual scales. As we have shown in both experiments, unmatchable surfaces can be a substantial portion of the sampled stimulus space even with displays as simple as the disk–ring arrangements used here. Thus, the extended scale provides a comparable measure for both increments and decrements and for surfaces that border with more than one surrounding region and are, therefore, an increment and a decrement at the same time. Second, the extended scale provides a more valid measurement of achromatic color for surfaces that are close to the unmatchable region and that would, therefore, be affected by floor or ceiling effects. As was shown in the second experiment, such effects are apparent with all the three kinds of scales employed here.

REFERENCES

Agostini, T., & Bruno, N. (1996). Lightness contrast in CRT and paper-and-illuminant displays. *Perception & Psychophysics*, **58**, 250-258.

Agostini, T., & Galmonte, A. (1999). Spatial articulation affects lightness. *Perception & Psychophysics*, **61**, 1345-1355.

Bonato, F., & Gilchrist, A. L. (1994). The perception of luminosity on different backgrounds and in different illuminations. *Perception*, **23**, 991-1006.

Bruno, N. (1992). Lightness, equivalent backgrounds, and the spatial integration of luminance. *Perception*, **21** (Suppl.), 80b.

Bruno, N., Bernardis, P., & Schirillo, J. (1997). Lightness, equivalent backgrounds, and anchoring. *Perception & Psychophysics*, **59**, 643-654.

Bruno, N., Jandó, G., Galmonte, A., & Agostini, T. (1998). Towards a psychophysics of lightness/brightness in disk-gradient-ring displays: Validation, slope, and size [Abstract]. *Investigative Ophthalmology & Visual Science*, **39**, S154.

Evans, R. M. (1959). Fluorescence and gray content of surface colors. *Journal of the Optical Society of America*, **49**, 1049-1059.

Evans, R. M. (1974). *The perception of color*. New York: Wiley.

Gilchrist, A. L. (1988). Lightness contrast and failures of constancy: A common explanation. *Perception & Psychophysics*, **43**, 415-424.

Gilchrist, A. L. (1994). Absolute versus relative theories of lightness perception. In A. L. Gilchrist (Ed.), *Lightness, brightness, and transparency* (pp. 1-33). Hillsdale, NJ: Erlbaum.

Gilchrist, A. L., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., Spehar, B., Annan, V., & Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, **106**, 795-834.

Heinemann, E. G. (1955). Simultaneous brightness induction as a function of inducing- and test-field luminances. *Journal of Experimental Psychology*, **50**, 89-96.

Heinemann, E. G. (1989). Brightness contrast, brightness constancy, and the ratio principle. *Perception & Psychophysics*, **45**, 89-91.

Jacobsen, A., & Gilchrist, A. (1988a). Hess and Pretori revisited: Resolution of some old contradictions. *Perception & Psychophysics*, **43**, 7-14.

Jacobsen, A., & Gilchrist, A. (1988b). The ratio principle holds over a million-to-one range of illumination. *Perception & Psychophysics*, **43**, 1-6.

Judd, D. (1966). Basic correlates of the visual stimulus. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 811-867). New York: Wiley.

Kardos, L. (1934). Ding und Schatten: Eine experimentelle Untersuchung über die Grundlage des Farbensehens [Thing and shadow: An experimental investigation concerning the basis of color vision]. *Zeitschrift für Psychologie*, **23**(Suppl.).

Katz, D. (1935). *The world of colour*. London: Kegan Paul, Trench, Trubner.

Land, E. H., & McCann, J. (1971). Lightness and retinex theory. *Journal of the Optical Society of America*, **61**, 1-11.

MacLeod, R. B. (1947). The effect of "artificial penumbrae" on the brightness of included areas. In *Miscellanea psychologica Albert Michotte* (pp. 138-154). Louvain: Institut Supérieur de Philosophie.

Schirillo, J., & Shevell, S. (1996). Brightness contrast from inhomogeneous surrounds. *Vision Research*, **36**, 1783-1796.

Wallach, H. (1948). Brightness constancy and the nature of achromatic colors. *Journal of Experimental Psychology*, **38**, 310-324.

Whittle, P. (1994). The psychophysics of contrast brightness. In A. L. Gilchrist (Ed.), *Lightness, brightness, and transparency* (pp. 35-110). Hillsdale, NJ: Erlbaum.

Whittle, P., & Challands, P. D. C. (1969). The effect of background luminance on the brightness of flashes. *Vision Research*, **9**, 1095-1110.

NOTE

1. We did not balance the split surround position (white always on the left). However, if there is any bias due to the position of the highest luminance in the global configuration, its effect is constant in all the conditions.

(Manuscript received April 25, 2001;
revision accepted for publication June 12, 2002.)