Judgments of average lightness and darkness: A further consideration of inverse attributes

DWIGHT W. CURTIS and STANLEY J. RULE
University of Alberta, Edmonton, Alberta T6G 2E9, Canada

Twenty subjects judged the average lightness and darkness of paired Munsell gray papers according to magnitude estimation and category rating instructions. The data from category rating were approximately consistent with an averaging model but those from magnitude estimation were not. An analysis in terms of a two-stage model suggested that category ratings and magnitude estimates of each attribute were produced on the basis of the same composition rule, but implicated different output transformations. In judging darkness, subjects appeared to have reversed the scales employed in judging lightness, by a linear transformation in the case of category rating and by a reciprocal transformation in the case of magnitude estimation. However, differences between the input parameters obtained from judgments of lightness and darkness suggested that the scales of these attributes constructed from judgments of average magnitude may be biased.

Judgments of inverse attributes, such as lightness and darkness, were first investigated by Torgerson (1960), whose principal finding was that scales of the two attributes constructed from magnitude estimates were related by a reciprocal function, whereas scales constructed from category ratings of the same stimuli were linearly related with a negative slope. Similar results have been obtained from judgments of roughness and smoothness (Stevens & Harris, 1962), loudness and softness (Schneider & Lane, 1963; Stevens & Guirao, 1962), and redness and paleness (Panek & Stevens, 1966).

Rule, Laye, and Curtis (1974) investigated the locus of the inverse relation between magnitude estimation scales of lightness and darkness in terms of a two-stage model of magnitude estimation (Attneave, 1962; Curtis, Attneave, & Harrington, 1968). The objective of the Rule et al. study was to determine whether the reciprocal form of the relation resulted because subjects responded in a similar way to two perceptual attributes that were reciprocally related or because a single continuum of perceived magnitude that was common to the two tasks was mapped onto numerical response scales that were reciprocally related. That is, we wished to determine whether the reciprocal form of the relation reflected a perceptual transformation or a response transformation.

According to the two-stage model, the relation between magnitude estimates and stimulus measures depends on a nonlinear response transformation as

well as an input, or sensory transformation, such that the empirical power function that typically describes the stimulus-response relation can be rewritten in the form,

$$Y = aX^{km}. \quad (1)$$

In Equation 1, Y denotes a numerical judgment of a stimulus with physical measure X , k and m are parameters of the input and output functions, respectively, and a is a coefficient that depends on the unit for X and Y . Given that k and m are both positive for magnitude estimation of lightness, the reciprocal relation between scales of lightness and darkness could stem either from a negative value of k , indicating that perceived darkness is a reciprocal function of perceived lightness, or from a negative value of m , indicating that perceptual magnitudes of the stimuli that were common to the two tasks were mapped onto numbers that were reciprocally related in the two tasks.

The usual magnitude estimation experiment in which subjects evaluate stimuli as they are presented one at a time does not provide a basis for discriminating between these two interpretations because the input and output transformations are confounded. However, Curtis et al. (1968) have shown that if the subjective difference between two stimuli is equal to the difference between their subjective magnitudes, the input and output transformations may be estimated from magnitude judgments of the differences between stimuli. Rule et al. (1974) found that the input functions from judgments of difference in lightness and darkness were well approximated by power functions with approximately equal positive exponents. This is consistent with the inter-

This research was supported by Grants A9582 and A0151 from the National Research Council of Canada. Requests for reprints should be sent to Dwight W. Curtis, Department of Psychology, Biological Sciences Building, University of Alberta, Edmonton, Alberta T6G 2E9, Canada.

pretation that perceptual magnitude is the same for both lightness and darkness. The output functions were also well described by power functions with exponents greater than unity, a result consistent with those of other studies in which the two-stage model was evaluated (e.g., Curtis, 1970; Curtis & Rule, 1972; Rule, Curtis, & Markley, 1970).

The present study differed from the Rule et al. (1974) investigation in that subjects estimated average lightness and darkness rather than difference in lightness and darkness and that category ratings as well as magnitude estimates were collected. It was conducted in an attempt to resolve two issues, one theoretical and the other methodological. The theoretical issue follows from an interpretation by Marks (1974) of the findings from some tests of the two-stage model, while the methodological issue stems from a difference between the results obtained by Rule et al. (1974) and those obtained by Curtis and Mullin (1975) from a reanalysis of data collected by Weiss (1972).

Marks (1974) questioned the assumption, essential to most tests of the two-stage model, that the subjective difference between stimuli is equivalent to the difference between the subjective magnitudes of the stimuli, arguing that perceived magnitude and perceived difference (or dissimilarity) may constitute different classes of sensory information. In particular, he suggested that there are two classes of perceived relations, one yielding magnitude scales and the other yielding dissimilarity scales. Magnitude scales are produced by ratio scaling procedures such as magnitude estimation, whereas dissimilarity scales are produced by interval judgment procedures, of which category rating is prototypical. He has included magnitude estimation of difference in the latter class. The essential difference between Marks' view and ours is that he considers the differences typically found between scales constructed from judgments of ratios and intervals to be perceptual in origin, whereas our two-stage model considers them to be due to differences in the mode of responding (Curtis & Rule, 1978). This disagreement is well illustrated by the way in which each position treats magnitude estimation of the differences between stimuli.

It is well established that when subjects evaluate differences between stimuli according to magnitude estimation instructions, the data obtained are not consistent with a simple linear (difference) model. However, a nonlinear (usually power) transformation can be found such that the properties of the transformed data approximate those of a difference model. That is, magnitude judgments of difference are monotonically related to perceived difference, but not necessarily linearly related. We have referred to this monotone relation as the output transformation,

on the assumption that it reflects a response bias characteristic of the magnitude estimation task. In contrast, Marks has suggested that the nonlinear component reflects a task-induced tendency on the part of subjects to characterize the dissimilarities among sensory stimuli in terms of the dissimilarities among numbers. From Marks' view, the central correlates of numerical responses are also classed as either dissimilarities or magnitudes, depending on the task required. In the usual magnitude estimation task, the subject is assumed to match the subjective magnitude of a number to the subjective magnitude of the stimulus where, for number, subjective magnitude is assumed to be linear with the response. He hypothesizes that subjective dissimilarity is related to subjective magnitude by a square-root function. Consequently, if subjects match number dissimilarity to stimulus dissimilarity when they are instructed to evaluate dissimilarity (difference) according to magnitude estimation instructions, the output function that results should be the inverse of the square-root dissimilarity function (i.e., the output function should be a power function with an exponent of 2.0).

A response scale based on the dissimilarity relation should have properties in common with category rating data and, as was noted earlier, category ratings of inverse attributes are linearly related with a negative slope. It follows that if subjects reverse their response scale in judging the inverse of an attribute, the reversed dissimilarity scale of number should be linear with the scale of the original attribute. Unfortunately, judgments of difference, such as those collected by Rule et al. (1974) do not provide a test of this prediction. Difference in darkness increases as stimulus difference increases, just as does difference in lightness, so that the response scales associated with the inverse attributes are positively correlated. In contrast, judgments of the average magnitude of two stimuli should yield negatively correlated scales for inverse attributes. Since, according to Marks, such scales are also based on dissimilarity information, they should provide a test of this prediction.

Some data collected by Weiss (1972) are relevant to this issue. He obtained both magnitude estimates and graphic ratings (a variant of category ratings) of the average darkness of Munsell grays. Curtis and Mullin (1975) reanalyzed these data, fitting to them a generalization of the two-stage model given by the expression,

$$Y_{ij} = a[wX_i^k + (1 - w)X_j^k]^m + b. \quad (2)$$

In Equation 2, Y represents a judgment of the average darkness of the stimulus pair with reflectances X_i and X_j , w and $1 - w$ are weights which sum to

unity for an averaging model, and k and m are the parameters of the input and output transformations, respectively. The additive constant b was required in fitting the model to the graphic ratings because these data define an equal interval scale at best. The constant was also included in the equation fitted to the magnitude estimation data on the assumption that it represents a constant error reflecting a displacement of the origin on the response scale. A detailed discussion of the constant is developed in Rule et al. (1974).

The input exponents yielded by the graphic rating and magnitude estimation data were both positive and were almost identical, consistent with the Rule et al. finding that values on the sensory scale increase as reflectance increases for darkness as well as for lightness. However, the output transformations differed. The graphic rating data yielded a positive exponent, close to unity, and a negative coefficient, as though subjects' response scale in judging darkness was a linear transform of their scale for average lightness. In contrast, magnitude estimates of average darkness yielded an output exponent of approximately -2 , suggesting both that subjects had employed a nonlinear response scale and that there had been a reciprocal mapping of average magnitude onto that scale. The evidence for a reciprocal response transformation is difficult to reconcile with Marks' interpretation that subjects were responding to the dissimilarity relations among numbers.

Although Weiss' data appear to be inconsistent with Marks' hypothesis, neither are they wholly consistent with the results obtained by Curtis and Rule (1972) and Rule et al. (1974) from judgments of difference. The input exponents of .22 yielded by Weiss' data were approximately two-thirds the value of the exponents obtained in the other experiments. However, Weiss' study differed from the Rule et al. study in several of its experimental features. In particular, the reflectance of the background on which Weiss presented his targets appears to have differed from those used in the other experiments. Weiss does not report this value, but his description of the experiment suggests that it was not uniform, and it seems quite possible that the conditions of contrast might have influenced the form of the psychophysical relation. Curtis and Rule (1972) obtained much smaller input exponents from magnitude judgments of difference in lightness when the targets were presented against a dark background than when the background was light.

The present study represented an attempt to resolve these issues. Subjects judged the average lightness and darkness of pairs of gray paper, as in Weiss' study, but the targets were presented against a high-reflectance background, as in the Rule et al. study. Several predictions are possible on the basis

of Marks' model and our two-stage model. First, both models predict the input exponent k (in Equation 2) to be invariant over judgment tasks, and its value should be equal to the value obtained by Rule et al. from magnitude estimates of difference in lightness and darkness. Second, both models predict a linear relation between category ratings of average lightness and darkness; the two values of m should be approximately equal, but the values of the coefficient, a , should differ in algebraic sign. Finally, the two models differ in their prediction of the output transformation for magnitude estimation of darkness. In this instance, Marks' model predicts a linear relation to hold between magnitude estimates of average darkness and magnitude estimates of average lightness; m should be approximately equal for the two cases, but the coefficient should differ in sign. The two-stage model predicts that the output function for average darkness should be the reciprocal of the output function for average lightness; m should be approximately equal in absolute value, but its sign should be positive for judgments of average lightness and negative for judgments of average darkness.

The empirical discrepancy between Weiss' and Rule et al.'s data raises the possibility that the assumption that k is invariant over tasks and attributes may be too strong. Several bases can be suggested for this discrepancy. The difference in value of the exponents may be attributable to a difference in background reflectance (in which case we would expect a high background reflectance to result in input exponents of around one-third for average darkness and lightness), to the requirement to evaluate averages rather than differences (in which we would expect input exponents of around .22 for average lightness and darkness), or to some asymmetry of the lightness and darkness continua for judgments of averages (in which case we would expect an input exponent of .22 for darkness and .33 for lightness).

METHOD

Twenty undergraduate and graduate students who served as paid subjects were presented with pairs of gray papers from the Munsell neutral value series. Seven stimuli were used with Munsell values ranging from 3.0 to 9.0 in increments of one unit. Their reflectances were 6.5%, 12.0%, 19.8%, 30.0%, 43.1%, 59.1%, and 78.7%. These were combined according to a 7 by 7 factorial design to provide 49 stimulus pairs. The stimulus pairs were viewed through 6.4-cm-diam circular openings in a 69.2 × 92.2 cm white surface, the reflectance of which was .954. The surface was indirectly illuminated by fluorescent tubes, such that the luminance of the background was 147 cd/m².

Each subject participated in four judgment tasks, magnitude estimation of average lightness and darkness and category rating of average lightness and darkness. The tasks were presented on successive days in a different random order for each subject. In the magnitude estimation conditions, the subjects were in-

structed to assign a number to the first pair which appeared to best represent its average lightness (darkness), and to assign numbers to subsequent pairs such that their values reflected their average magnitude relative to the average magnitudes of other pairs. In the category rating tasks, the subjects were instructed to rate the average lightness (darkness) of the pairs on a scale of 1 to 11, such that a given category difference reflected about the same amount of difference in lightness (darkness). Two papers whose Munsell values were 3 and 9 were presented first to enable the subject to anchor his scale.

Ten practice trials were presented in each task before the collection of data began, and the subject was then questioned to ascertain his understanding of the instructions. Each session consisted of the 10 practice trials, followed by two replications of the stimulus series.

RESULTS

The average judgments, pooled over replications and subjects are plotted in Figure 1 as a function of the Munsell value of the stimulus on the right. The lines connect points associated with a constant value for the left target. Medians were used in pooling category ratings and geometric means were used in pooling magnitude estimates.

If a linear (simple averaging) model held, consistent with instructions, the data should plot as a set of parallel functions. This appears to be the case for the category ratings of lightness and darkness, but not for the magnitude estimates of the same attributes. For the category ratings of lightness, parallelism was confirmed by an analysis of variance; the interaction between left and right targets was not significant [$F(36,684) = 1.11$]. The category ratings of darkness exhibited a slight, but significant, interaction [$F(36,684) = 1.47, p < .05$], the source of which is not evident in Figure 1. In contrast, there is clear evidence of interactions for both sets of magnitude estimation data. The F ratios for the interaction were: $F(36,684) = 2.38, p < .001$, for the judgments of lightness and $F(36,684) = 2.11, p < .001$, for the judgment of darkness.

The foregoing results suggest that the linear model adequately describes category ratings of average lightness and possibly category ratings of average darkness; however, such a model does not appear to describe either set of magnitude estimation data. Interactions such as these may indicate either that

Figure 1.

the composition rule employed in these tasks was not a simple averaging rule, or that the response scale that was employed was not a linear representation of judgments. Curtis and Mullin (1975) considered the latter possibility, finding that Weiss' data from magnitude estimation and category rating of average darkness yielded different output transformations for the two tasks, but once these nonlinear components were accounted for the data were consistent with a single composition rule and psychophysical relation.

Equation 2 was fitted to the data from the four judgment tasks by the Gauss-Newton method (Hartley, 1961), an iterative procedure for estimating the parameters of nonlinear equations which provide a best-fit in terms of a least-squares criterion. The resulting equations are given in Table 1.

DISCUSSION

Several features of the results are to be noted. First, the equations describing magnitude estimation and category rating of average lightness suggest that the same composition rule was applied in the two tasks, but that performances differed with respect to their output transformations. Values of the input exponent k were virtually identical for the two sets of data, and estimates of w indicate that the right and left stimuli were given about equal weight. However, the values of the output exponent m are quite different. Its value of 1.07 for the category ratings of average lightness implies a close-to-linear relation between average lightness and responses, but the exponent of 2.28 from magnitude estimation indicates a highly nonlinear output transformation. These results are similar to those from other experiments employing the two-stage model.

The primary difference between category ratings and magnitude estimates of average darkness also appears to be associated with the respective output transformations; however, here the difference is somewhat more complex. For the category ratings, the output exponent was positive and the coefficient was negative, as though subjects in representing average darkness had simply reversed the ends of the scale used in evaluating average lightness (however, note that the output exponent of 1.3 implies a moderately nonlinear output transformation, which may have contributed to the significant interaction in the analysis of variance). Although the output

exponent of -2.06 from magnitude estimation is comparable in absolute magnitude to the similar exponent from magnitude estimation of average lightness, its negative sign implies two sources of nonlinearity, a component that is presumably a consequence of the use of a nonlinear response scale, and a second component resulting from a reciprocal assignment of values from that scale to stimuli. Either source alone might be expected to result in a significant interaction in the analysis of variance. This result is essentially the same as that obtained by Curtis and Mullin (1975) from Weiss' magnitude estimates of average darkness.

Finally, the input exponents from judgments of average darkness (.17 and .22) were much smaller than the corresponding exponents from judgments of average lightness (.42 and .43). Those from judgments of average darkness are comparable to the estimates of .21 and .22 from Weiss' data, indicating that the latter results were a reliable consequence of judging average darkness rather than an artifact of some feature of Weiss' experiment, such as the contrast associated with the reflectance of the background.

In the Rule et al. study, the input exponent estimated from judgments of difference in lightness did not differ appreciably from the exponent estimated from judgments of difference in darkness. This finding, which implies that a single sensory scale underlies judgments of lightness and darkness, contrasts with the present finding that the relation between the scales for lightness and darkness is nonlinear. It seems likely that this difference between results from the two classes of judgment task is due to biases in judging averages which operate in opposite directions depending on whether the attribute evaluated is lightness or darkness. It may be noted that the mean of the exponents from the two averaging tasks of .31 is almost identical to the mean exponent from the data of Rule et al.

It is not clear whether this bias (if it is a bias) is characteristic of judgments of averages in general, or whether it occurs only on the lightness and darkness continua. Although direct evidence from such continua as loudness and softness of tones or length and shortness of durations are not available, some indirect evidence is available which suggests that it is not the general case. For instance, judgments of difference in duration and average duration (Curtis & Rule, 1977) and difference in loudness and average

Table 1
Best-Fitting Expressions (Equation 2) Describing the Relation Between Stimulus
and Response for the Data of Each of the Four Conditions

	Lightness	Darkness
Category Rating	$Y_{ij} = 1.89(.50X_i^{.43} + .50X_j^{.43})^{1.07} - 3.35$	$Y_{ij} = 20.56 - 5.70(.49X_i^{.22} + .51X_j^{.22})^{1.30}$
Magnitude Estimation	$Y_{ij} = .71(.49X_i^{.42} + .51X_j^{.42})^{2.28} - 2.32$	$Y_{ij} = 197.71(.52X_i^{.17} + .48X_j^{.17})^{-2.05} - 36.89$

loudness (Curtis & Mullin, 1975; Mullin & Curtis, 1973, Note 1) do not appear to yield differing values of the input exponent as is the case for the present results.

The reason why bias might be introduced in judgments of average lightness and darkness is not clear. Possibly, the bias might be a consequence of properties of these sensory continua that make them unique relative to other prothetic continua. If illuminance is held constant, the physical continuum underlying the perception of lightness and darkness is bounded at one extreme by 0% reflectance and at the other by 100% reflectance. Unlike other prothetic continua, subjects find it as easy to judge relative darkness as to judge relative lightness, probably because an origin is defined by the respective limiting values on the physical scale. But each of the inverse attributes has an upper limit, as well as a lower limit, and it may be the case that the constraints imposed upon judgments of averages may tend to result in floor or ceiling effects that bias the data in different directions depending on whether the subject is evaluating lightness or darkness.

Despite the foregoing qualifications, the present results, together with the results of Curtis and Mullin's reanalysis of Weiss' data, tend to diminish the credibility of Marks' hypothesis that subjects attend only to the dissimilarity relations among numbers when they evaluate stimulus pairs according to a "difference" composition rule. If such were the case, the appropriate reciprocity between response scales for the inverse attributes should be consistent with a linear function. This appears to be approximately the case for category ratings, but not for magnitude estimations. In the latter case, values of m should have been approximately equal to (positive) 2.0 for magnitude estimation of both average lightness and average darkness, a value that was reasonably well approximated for the former data. However, the least squares solution yielded a value closer to -2.0 in the latter case, indicating that a reciprocal relation rather than a linear relation holds between the response scales. Nevertheless, as a final check for the inverse attributes, Equation 2 was fitted to the data from magnitude estimation of darkness with m held constant at 2.0. This solution yielded a least squares estimate of the input exponent k of $-.45$ and weights associated with the left and right targets of $.2$ and $.8$, respectively. This estimate of the input exponent was quite different from the one from category judgments of average darkness, and there seemed to be little justification for weighting the right stimulus four times as heavily as the left stimulus. An additional analysis held the weights w constant at $.5$ and held m constant at 2, yielding an estimate of the input exponent k of $-.01$, a value unlike any other that have been reported for this continuum. In both of the foregoing analyses, a sizable nonlinear residual component

remained in the relation between the data and values predicted by the model.

REFERENCE NOTE

1. Mullin, L. C., & Curtis, D. W. *Simple and disjunctive reaction time as a function of loudness*. Paper presented at the meeting of the Western Psychological Association, Anaheim, California, 1973.

REFERENCES

- ATTNEAVE, F. Perception and related areas. In S. Koch (Eds.), *Psychology: A study of a science*. New York: McGraw-Hill, 1962.
- CURTIS, D. W. Magnitude estimations and category judgments of brightness and brightness intervals: A two-stage interpretation. *Journal of Experimental Psychology*, 1970, **83**, 201-208.
- CURTIS, D. W., ATTNEAVE, F., & HARRINGTON, T. A test of a two-stage model of magnitude estimation. *Perception & Psychophysics*, 1968, **3**, 24-31.
- CURTIS, D. W., & MULLIN, L. C. Judgments of average magnitude: Analyses in terms of the functional measurement and two-stage models. *Perception & Psychophysics*, 1975, **18**, 299-308.
- CURTIS, D. W., & RULE, S. J. Magnitude judgments of brightness and brightness difference as a function of background reflectance. *Journal of Experimental Psychology*, 1972, **95**, 215-222.
- CURTIS, D. W., & RULE, S. J. Judgments of duration relations: Simultaneous and successive presentation. *Perception & Psychophysics*, 1977, **22**, 578-584.
- CURTIS, D. W., & RULE, S. J. The binocular processing of brightness information: A vector-sum model. *Journal of Experimental Psychology: Perception and Performance*, 1978, **4**, 132-144.
- HARTLEY, H. O. The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, 1961, **3**, 269-280.
- MARKS, L. E. On scales of sensation: Prolegomena to any future psychophysics that will be able to come forth as science. *Perception & Psychophysics*, 1974, **16**, 358-376.
- PANEK, D. W., & STEVENS, S. S. Saturation of red: A prothetic continuum. *Perception & Psychophysics*, 1966, **1**, 59-66.
- RULE, S. J., CURTIS, D. W., & MARKLEY, R. P. Input and output transformations from magnitude estimation. *Journal of Experimental Psychology*, 1970, **86**, 343-349.
- RULE, S. J., LAYE, R. C., & CURTIS, D. W. Magnitude judgments and difference judgments of lightness and darkness: A two-stage analysis. *Journal of Experimental Psychology*, 1974, **103**, 1108-1114.
- SCHNEIDER, B., & LANE, H. Ratio scales, category scales, and variability in the production of loudness and softness. *Journal of the Acoustical Society of America*, 1963, **35**, 1953-1961.
- SCHNEIDER, B., PARKER, S., FARRELL, G., & KANOW, G. The perceptual basis of loudness ratio judgments. *Perception & Psychophysics*, 1976, **19**, 309-320.
- STEVENS, S. S., & GUIRAO, M. Loudness, reciprocity, and partition scales. *Journal of the Acoustical Society of America*, 1962, **34**, 1466-1471.
- STEVENS, S. S., & HARRIS, J. R. The scaling of subjective roughness and smoothness. *Journal of Experimental Psychology*, 1962, **64**, 489-494.
- TORGERSON, W. S. Quantitative judgment scales. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications*. New York: Wiley, 1960.
- WEISS, D. J. Averaging: An empirical validity criterion for magnitude estimation. *Perception & Psychophysics*, 1972, **12**, 385-388.