Perception & Psychophysics
1978, Vol. 24 (3), 243-252

A theory of relative judgment

ROBERT F. FAGOT
University of Oregon, Eugene, Oregon 97403

This paper is concerned with the scaling method of "ratio estimation." The simple theory that equates reported ratio judgments to ratios of psychological magnitudes is first considered, then two close relatives of this theory are formulated, each of which places weaker constraints on the structure of the data. Structural conditions are stated that express the relations that must hold among observed ratio judgments for each of the models. The models proposed are "cumulative" in the sense that the second is a weakened version of the first, and the third a weakened version of the second. A special feature of the models is that they may be tested entirely in terms of observables, avoiding the necessity of scale construction prior to testing. Tests were carried out on data from 9 published studies. The strongest model, typically used in scale construction using ratio estimation data, was generally inadequate, showing large systematic errors. However, the weakest version generally passed the tests of internal consistency, and the model equation provided a basis for constructing ratio scales utilizing bias parameters.

This paper is concerned with the scaling method of *ratio estimation*. In a typical task, a pair of stimuli (a, b) is presented, and the subject is instructed to respond with a number that corresponds to the sensation "ratio" of a to b relative to a defined attribute. In another variation, called "free ratio estimation" (Mashhour, 1964), subjects assign a number to each of the two stimuli such that the ratio of the numbers assigned reflects the sensation ratio. Thus the method applies to any response that can be transformed to a ratio. The method does not require a modulus, and usually all possible pairs are presented.

The method of ratio estimation has a long history (Ekman, 1958; Stevens, 1958), but recently the simple theory that equates the reported sensation ratio to a ratio of psychological magnitudes has been questioned (e.g., Fagot, Stewart, & Kleinknecht, 1975; Sjöberg, 1971). If the simple theory does not lead to valid ratio scales, how might the theory be modified without abandoning the basic assumption that subjects are capable of making ratio judgments that are consistent in a well-defined sense?

This paper answers the question by developing two close relatives of the simple theory of ratio estimation that place weaker constraints on the structure of the data. The different versions constitute a theory of *relative judgment* in the obvious sense that a stimulus is judged relative to another stimulus. This relativity is in no sense trivial, since it leads to the formulation of tests of internal consistency not possible with methods such as the pure form of magnitude estimation in which stimuli are judged one at a time.

Work on this paper was carried out, in part, during the author's tenure as a Fellow at the Netherlands Institute for Advanced Study, Wassenaar, Holland.

The basic question asked is: Assuming that a ratio scale exists, what relations must hold among observed ratio estimations? The answer depends, of course, on the model; i.e., the assumed relationship between ratio estimations and psychological magnitudes. Three models will be presented, and conditions placing constraints on the observables will be derived from the models. The models permit tests to be carried out entirely in terms of observables, avoiding the necessity of scale construction prior to testing. Tests will be carried out on data from 9 published studies.

THEORY

The reported ratio estimation of stimulus a to stimulus b relative to a defined attribute will be denoted by R_{ab}. Ψ will denote the *scaling function* that assigns real numbers representing the psychological magnitudes of stimuli a, b, c, ... For convenience, the alphabetical ordering of the letters used to designate the stimuli will reflect the subjective ordering of the stimuli; i.e., stimulus a is less than stimulus b in the subjective ordering, etc. The set of stimuli will be designated S.

The first model is expressed by the following simple equation:

$$R_{ab} = \Psi_a/\Psi_b. \qquad (1)$$

Equation 1 will be referred to as the *classical* (C) model (Fagot, Stewart, & Kleinknecht, 1975). The C model is part of the lore of ratio scaling, and its use persists in spite of numerous disconfirmation studies (see Results section). It is included primarily as a point of reference in considering possible weakened versions.

Because of the poor fit of Equation 1 to ratio estimation data, Eisler (1960) proposed the following modification:

$$R_{ab} = \beta\Psi_a/\Psi_b, \tag{2}$$

where β may be interpreted as a *bias parameter* that may account for the interactive effect of judging pairs of stimuli. Equation 2 will be referred to as the *constant bias* (CB) model.

The final model is the *relative bias* (RB) model:

$$R_{ab} = \beta_b\Psi_a/\Psi_b. \tag{3}$$

The RB model permits bias to vary depending on the pair of stimuli judged. Such a model could, for example, account for contrast and assimilation effects, with the magnitude of the bias depending on the larger stimulus.

For all three models, Ψ is assumed to be a ratio scale, and structural conditions (testable consequences) for each model—expressed as relations among observed ratio estimations R_{ab}—will be derived. Each condition will be a testable consequence of one or more models, and tests of the conditions will provide a basis for a choice among the three models. For illustrative purposes, each condition will be tested against the numerical example of an error-free "ratio matrix" for four stimuli presented in Table 1.

The weakest condition constraining the R_{ab} is C0 —*Monotonicity of* R_{ab}.

Rather than state the condition formally, it is illustrated in Table 1. C0 states that if the stimuli are ordered as given in Table 1, then the R_{ab} must decrease from left to right for each row ($R_{ab} > R_{ac} > R_{ad}$; $R_{bc} > R_{bd}$), and increase from top to bottom for each column ($R_{ac} < R_{bc}$; $R_{ad} < R_{bd} < R_{cd}$). Inspection shows that monotonicity is satisfied in Table 1. Monotonicity is a necessary condition for each of the models. Serious failure of C0 would be discouraging for any attempt to scale based on ratio estimation.

The following three conditions place stronger constraints on the data.

Condition 1 (C1)—Ratio consistency: For all a, b, c in S,

$$R_{ac} = R_{ab}R_{bc}. \tag{4}$$

Table 1
Hypothetical Ratio Matrix of R_{ab}

Stimuli	b	c	d
a	.80	.30	.20
b		.75	.50
c			.80

Equation 4 follows directly from Equation 1, but not from Equations 2 or 3. It is therefore a necessary condition for the C model but not for the CB or RB model. Condition 1 states essentially that ratio estimations behave like numerical ratios.

Note that the hypothetical data of Table 1, although satisfying C0, do not satisfy C1. For example, $R_{ac} = .30$, but $R_{ab}R_{bc} = (.80)(.75) = .60$.

It is easy to show from Equation 2 that for any triple (a, b, d),

$$\beta = R_{ab}R_{bd}/R_{ad}, \tag{5}$$

and hence the bias parameter is expressed entirely in terms of observables. Since for another triple (a, c, d), $\beta = R_{ac}R_{cd}/R_{ad}$, we come immediately to the principal testable consequence of the CB model:

Condition 2 (C2)—Product constancy: For all a, b, c, d in S,

$$R_{ab}R_{bd} = R_{ac}R_{cd}. \tag{6}$$

Condition 2 states that for a pair of "outer" stimuli (a, d), the product $R_{ax}R_{xd}$ is *constant* under changes in the "inner" stimulus x. Note that C2 is "weaker" than C1, since C1 adds the further condition that $R_{ax}R_{xd} = R_{ad}$, i.e., that R_{ad} must be that constant. Clearly, C1 entails C2, but C2 does not entail C1. Hence, C1 places stronger constraints on the structure of the data than does C2, and there may exist sets of data satisfying C2 but not C1.

Referring again to the hypothetical data of Table 1, we note that $R_{ab}R_{bd} = (.80)(.50) = .40$ and $R_{ac}R_{cd} = (.30)(.8) = .24$, and therefore C2 is not satisfied.

Equation 5 also provides a means of formulating the CB *parametric form* of C1:

$$R_{ac} = (1/\beta)R_{ab}R_{bc}, \tag{7}$$

which indicates the way in which the CB model may correct for possible systematic error in C1.

Now consider the RB model: Writing the algebraic identity $(\Psi_a/\Psi_c) = (\Psi_a/\Psi_b)(\Psi_b/\Psi_c)$ and substituting from Equation 3, it follows that for any triple (a, b, c)

$$\beta_b = R_{ab}R_{bc}/R_{ac}. \tag{8}$$

Thus, for the RB model, the bias parameter is defined for the inner stimulus of each triple, and for n stimuli there are $(n-2)$ bias parameters with β_1 and β_n not defined. The latter restriction presents no difficulty in testing the model which is carried out entirely in terms of observables.

From Equation 8 it follows that

$$\beta_b = (R_{ab}R_{bd})/R_{ad} = (R_{ab}R_{bc})/R_{ac}, \tag{9}$$

from which Condition 3 follows directly:

Condition 3 (C3)—Ratio constancy: For all a, b, c, d in S,

$$(a)\ R_{ad}/R_{bd} = R_{ac}/R_{bc},$$

or equivalently,

$$(b)\ R_{ac}/R_{ad} = R_{bc}/R_{bd}. \qquad (10)$$

The two forms of C3 state that the ratios R_{ax}/R_{bx} (a) and R_{xc}/R_{xd} (b) are *constant* under changes in the stimulus x.

Note that if β is substituted for β_b in Equation 9, then the revised equations follow from Equation 5 and also entail Equation 10, showing that *C3 is a necessary condition for the CB as well as the RB model*. Hence, C2 *and* C3 are testable consequences of the CB model.

Referring to the hypothetical data of Table 1, we find that C3 is satisfied: $R_{ad}/R_{bd} = R_{ac}/R_{bc} = 2/5$. Hence only the RB model perfectly satisfies the ratio estimations of Table 1.

Equation 8 shows how to formulate the RB parametric form of C2:

$$(1/\beta_b)R_{ab}R_{bd} = (1/\beta_c)R_{ac}R_{cd}, \qquad (11)$$

which indicates the way in which the RB model may correct for systematic errors in C2.

The relations among the three structural conditions are as follows. First, C1 entails C2 and C3, but neither C2 nor C3 (singly or in combination) entails C1. Thus C1 is seen to be a very powerful condition. Second, C2 and C3 are independent, and sets of data may exist satisfying C2 but not C3, and vice versa. For example, Table 1 satisfies C3 but not C2; and if R_{ac} is changed to $R_{ac} = .50$, then C2 would be satisfied but C3 violated.

A characterization of the models is summarized in Table 2. The models are *cumulative*: Starting with the weakest model (RB) (i.e., weakest in the sense that it places the weakest constraints on the ratio estimations), successive strengthening is accomplished by adding on conditions one at a time, such that the conditions for a weaker model are a subset of the conditions for a stronger model. Of course, C2 and C3 are redundant for the C model since they are entailed by C1. The cumulative nature of the models is a nice property from the point of view of empirical tests and scale construction. Given that C0 is satisfied: (1) If C1 is satisfied, further testing is unnecessary and Equation 1 can be used to construct a scale. (2) If C1 is not satisfied, but C2 and C3 are satisfied, then Equation 2 can be used to estimate β and construct a scale. (3) If C1 and C2 are not satisfied, but C3 is satisfied, then Equation 3 can be used to estimate parameters β_b (b = 2, 3, . . . , n − 1) and construct a scale.

An interesting connection of relative judgment theory to factor analysis can be demonstrated. If we treat Equations 1, 2, and 3 as "fundamental equations" of factor analysis, without unique factors (Mulaik, 1972, p. 100), then C1, C2, and C3 can be generated by assuming a single factor (Ψ_a/Ψ_b). The ordinary symmetric correlation matrix is replaced by a ratio matrix that is not symmetric. The ratio matrices of the three models differ with respect to diagonal elements (R_{aa}) and elements below the main diagonal (R_{ba}). For the C model, $R_{aa} = 1$ and $R_{ba} = 1/R_{ab}$. For the CB model, $R_{aa} = \beta$ and $R_{ba} = \beta^2/R_{ab}$. For the RB model, $R_{aa} = \beta_a$ and $R_{ba} = \beta_a\beta_b/R_{ab}$.

Assuming a single factor, the rank of the ratio matrix must be one; i.e., all determinants of order two must vanish. Solving the resulting determinant equations (with appropriate R_{aa} and R_{ba}, depending on the model) generates the structural conditions, consistent with the characterization of the models in Table 2. An interesting finding is the fact that C3 expresses for ratios what the Spearman tetrad-difference criterion expresses for correlations.

Related Theory

Equation 1, the C model, has been assumed by most investigators using ratio estimation, and it corresponds to Sjöberg's (1971) Model 1. Condition 1 has been formulated and tested (but not confirmed) by Eisler (1960), Fagot and Stewart (1969), and Goude (1962). Equation 2, the CB model, was proposed by Eisler (1960) and Goude (1962) and corresponds to Sjöberg's (1971) Model 2. Sjöberg (1971) also proposed a variable/standard (VS) model that assigned a different scale value to a stimulus depending on its status as comparison (variable) or standard stimulus. He further pointed out that the VS model

Table 2
Characterization of Models

Model	Structural Conditions (Testable Consequences)	Representation	Bias Parameter $R_{ab}R_{bc}/R_{ac} =$
Classical (C)	C0, C1, (C2), (C3)	$R_{ab} = \Psi_a/\Psi_b$ (Equation 1)	1
Constant Bias (CB)	C0, C2, C3	$R_{ab} = \beta\Psi_a/\Psi_b$ (Equation 2)	β
Relative Bias (RB)	C0, C3	$R_{ab} = \beta_b\Psi_a/\Psi_b$ (Equation 3)	β_b

was indistinguishable from a model of the form

$$R_{ab} = (\beta_a\beta_b)(\Psi_a/\Psi_b). \quad (12)$$

It can be shown that Condition 3, although not formulated by Sjöberg, follows directly from Equation 12, which therefore has the same directly testable consequences as the RB model.

Fagot, Stewart, and Kleinknecht (1975) formulated a model incorporating a single bias parameter to account for both interval and ratio judgments. However, that model is more complicated than those considered in this paper, and conditions expressed entirely in terms of observables were not presented. Until such conditions are formulated, a comparison with the present theory is difficult.

Krantz (1972) developed a qualitative theory based on certain empirical generalizations, mainly magnitude estimation, cross-modality matching, and pair consistency [called ratio consistency (C1) in this paper]. Condition i is the main point of contact to the present theory, but a point of view shared with Krantz, and with R. N. Shepard as reported by Krantz, is that subjects judge *pairs* of stimuli, not single stimuli. This approach is embodied in the notion of *relative judgment* advanced in this paper.

The main theoretical contributions of the present theory are Conditions 2 and 3, which have not thus far been stated or tested, and the implications of the three conditions for scale construction and testing.

RESULTS AND DISCUSSION

Monotonicity will be tested first, then Conditions 1-3 will be analyzed in two ways. Each condition will be tested in the form Y = X:

$$C1: Y = R_{ac}, X = R_{ab}R_{bc}$$

$$C2: Y = R_{ac}R_{cd}, X = R_{ab}R_{bd}$$

$$C3: Y = R_{ad}R_{bc}, X = R_{ac}R_{bd}.$$

C3 is expressed above in its product form to be consistent with the forms of C1 and C2.

The two ways in which the conditions will be analyzed are: (1) tests of *agreement*, in the sense of *reliability*, between Y and X; and (2) tests for the presence of *systematic errors* in the plot of Y as a function of X.

The conditions will be tested against 9 published studies, the only studies found by the author that presented the ratio matrices necessary for testing the conditions. The studies are summarized in the first three columns of Table 3.[1] Study No. 4 combines results for 15 individual subjects listed in Table 4.

Test of Monotonicity

For 7 of the 9 studies, there were no violations of monotonicity. Study No. 4 had 19 violations out of 300 (6.3%) instances (pairs of ratios). However, the magnitudes of the reversals were small, with a median of .016. The violations were restricted to 10 of the 15 subjects.

The nature of the violations was interesting—all were *row* violations. Since, in this study, the fixed standard was always the brighter member of the pair, the violation instances show that the brightness judgments of a single comparator relative to a set of standards did not satisfy monotonicity, whereas judgments of several comparators relative to a single standard did without exception satisfy monotonicity.

The second study showing violations of monotonicity was Study No. 10 (odor intensity), which had six violations out of 40 instances (15%), with a median reversal of .035. Five of the six violations were *column* violations as contrasted to the result in Study No. 4. It will be shown below that the data of this study fit the conditions least well of all the studies (in the sense of reliability), although systematic errors were not detected. Cases violating monotonicity were not deleted in the reliability and systematic errors analysis.

Reliability

A one-way repeated measures analysis of variance

Table 3
Intraclass Correlation Coefficients (ICC) and Tests for Systematic Errors (TSE)

Study	Attribute	Reference	ICC			TSE		
			C1	C2	C3	C1	C2	C3
1	Size of Circular Surfaces	Ekman (1958)	.994	.981	.981	*	*	*
2	Angles	Goude (1962)	.981	.991	.996	††	**	*
3	Brightness	Fagot, Stewart, & Kleinknecht (1975)	.943	.923	.982	††	††	††
4	Velocity	Mashhour (Note 1)	.928	.976	.980	†	*	*
5	Heaviness	Eisler (1960)	.893	.982	.981	††	**	*
6	Darkness	Ekman, Goude, & Waern (1961)	.863	.983	.974	††	**	††
7	Weight	Goude (1962)	.837	.978	.987	**	*	*
8	Area	Ekman, Goude, & Waern (1961)	.781	.972	.901	**	*	*
9	Odor Intensity	Engen & Lindström (1963)	.691	.308	.811	*	*	*

*Nonsignificant **$p < .05$ †$p < .01$ ††$p < .001$

Table 4
Intraclass Correlation Coefficient (ICC) and Tests for Systematic Errors (TSE): Individual Subjects (Fagot, Stewart, & Kleinknecht, 1975)

	ICC			TSE		
Subject	C1	C2	C3	C1	C2	C3
K.H.	.991	.987	.989	*	*	*
J.B.	.994	.977	.986	*	*	*
D.S.	.993	.979	.991	*	*	*
D.K.	.983	.969	.987	*	*	*
T.M.	.976	.935	.978	†	*	*
R.H.	.972	.967	.996	††	**	*
T.T.	.956	.963	.994	††	**	*
D.R.	.950	.929	.971	†	*	*
D.L.	.948	.954	.987	**	**	*
D.S.R.	.944	.916	.993	*	*	*
L.C.	.932	.958	.970	†	*	*
W.H.	.912	.909	.984	**	**	**
G.S.	.911	.898	.972	†	*	**
L.B.	.906	.926	.962	†	**	*
M.W.	.905	.703	.984	**	*	†

*Nonsignificant **p < .05 †p < .01 ††p < .001

design was used to obtain reliability measures—the intraclass correlation coefficient (ICC)—for each condition (Bartko, 1976). Relative to an ordinary repeated measures design, Y and X correspond to two "ratings" and "stimulus sets" replace "subjects." For example, for C1, the stimulus sets are the triples of stimuli, each of which provides a datum for testing the condition. For C2 and C3, the stimulus sets are tetrads.[2] The closer the agreement between Y and X, the higher the reliability.

Referring back to Tables 3 and 4, the ICCs for each condition are presented in Columns 4-6. The studies (Table 3) and subjects (Table 4) are ordered from highest to lowest with respect to the ICC for C1. Theoretically, the ICC parameter values for C2 and C3 should be higher than for C1, but the estimators in the tables may not be, and are not in all cases.

What we want to ask of these data is whether the reliabilities of the three conditions are satisfactorily high, whether C2 and C3 show improvement over C1, and whether there are systematic trends in reliability coefficients. Clearly, if the base-rate reliability of C1 is very high (as is the case for the first two studies in Table 3), then improvement in C2 and C3 cannot be expected. However, it can be seen that the reliability of C1, though high for the first few studies in Table 3, decreases markedly through No. 9; but except for outlier No. 9, C2 and C3 maintain relatively high reliability in the face of steadily decreasing reliability for C1. In Table 4, the reliabilities for individual subjects tend to be higher for C1, but the same trend is present for C3 relative to C1, but not for C2 relative to C1.

This kind of analysis suggests a regression approach. For example, if we regress ICC(C3) on ICC(C1), do we get a small slope with high intercept, entailing high and relatively constant reliability for C3, little influenced by changes in reliability for C1?

A graphical representation is given in Figure 1, which shows four plots (A) ICC(C3) as a function of ICC(C1) for the 9 studies in Table 3; (B) ICC(C3) as a function of ICC(C1) for the 15 subjects in Table 4; (C) ICC(C3) as a function of ICC(C2) for Table 4; and (D) ICC(C2) as a function of ICC(C1) for eight studies from Table 3 (outlier No. 9 deleted).

The short-dashed line in each plot is the identity function. To sustain a hypothesis of equal reliability for the two conditions, all data points should cluster randomly about the line, clearly not the case for any of the plots. The long-dashed line in each plot is the best-fitting straight line. Two lines are plotted in

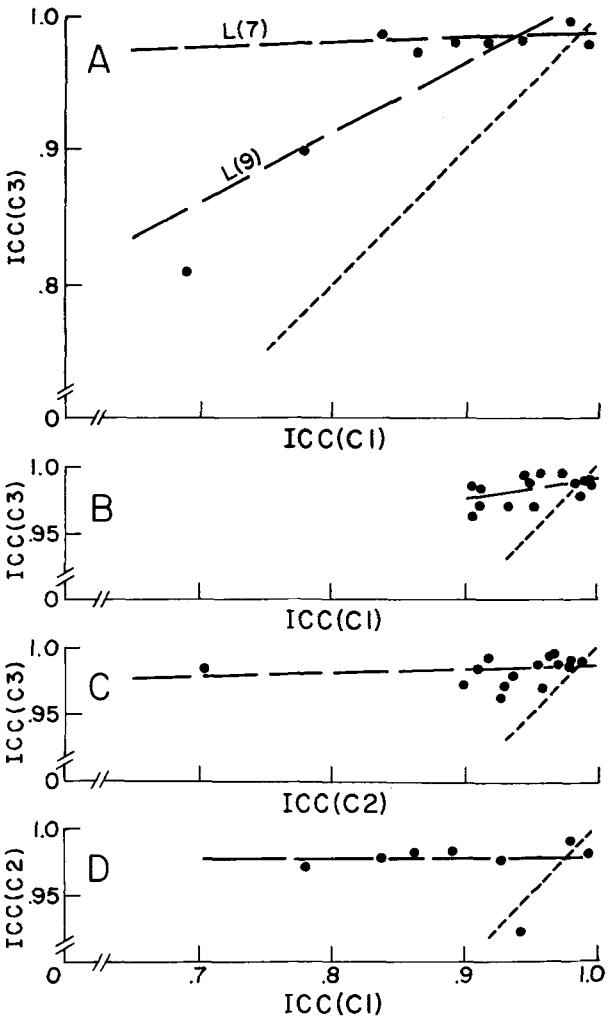

Figure 1. Comparisons of conditions via plots of intraclass correlation coefficients (ICC). Panel A: 9 studies from Table 3. Panels B and C: 15 subjects from Table 4. Panel D: 8 studies from Table 3 (outlier No. 9 deleted). In each panel, the short-dashed line is the identity line and the long-dashed line is the best-fitting line. In Panel A, L(9) is the best-fitting line for all 9 studies, and L(7) for 7 studies with two outliers deleted.

Table 5
Reliability Slopes and Intercepts

Regression of ICC(C_j) on ICC(C_i)($j > i$)		Slope	Intercept
C3 on C1	L9 (Panel A)	.533**	.487
	L7 (Panel A)	.040	.946
	15 subjects (Panel B)	.167*	.824
C3 on C2	9 studies (Table 3)	.242**	.737
	15 subjects (Panel C)	.023	.961
C2 on C1	9 studies (Table 3)	1.614**	−.519
	8 studies (Panel D)	−.014	.986
	15 subjects (Table 4)	1.395**	−.396

**Hypothesis of zero slope rejected: $p < .05$.*
***Hypothesis of zero slope rejected: $p < .01$.*

Panel A L(9) for all 9 studies and L(7) for the two outliers deleted.

Table 5 gives estimates of slopes and intercepts. Note that the hypothesis of zero slope for the regression lines in Figure 1 was rejected only for Panel B and L(9), Panel A. The intercepts give the estimated *lower bound* reliabilities: of C3 for zero reliability of C2 or C1, and of C2 for zero reliability of C1.

Comparison of C3 with C2 and C1 by means of Figure 1 and Table 5 indicates that, in general, C3 maintains a relatively constant and high reliability in the face of steadily decreasing reliabilities for C2 and C1. In addition, extrapolation of these data suggests that reasonable reliabilities would be maintained by C3 in the face of very bad fits to C1 and C2.

Comparison of C2 and C1 gives a mixed message. The data of Table 4 for individual subjects do not show higher reliabilities for C2, but if one outlier is omitted from Table 3, then Table 5 and Panel D show a near zero slope for the regression of C2 on C1 and a very high lower bound reliability for C2.

Systematic Errors

The analysis in this section will consider (1) statistical tests for the presence of systematic errors, and (2) model effects on the direction and magnitude of systematic errors and on scale values.

The ICC provides a measure of the agreement between Y and X but is not sensitive to error direction, i.e., to the presence of systematic errors in the plot of Y as a function of X (e.g., a possible tendency toward $X = R_{ab}R_{bc} > Y = R_{ac}$). In order to test for the presence of systematic errors, simple t tests for paired differences were carried out. Using the notation introduced at the beginning of this section, each null hypothesis is of the form $Y = X$. Then the data for the test consist of sets of paired differences $\hat{Y}_i - \hat{X}_i$.

Referring back to Tables 3 and 4, results of the statistical tests are shown in the last three columns. Note the relatively high reliabilities for many of the cases of statistically significant results, demonstrating the lack of sensitivity of the ICC to the presence of systematic errors. Reliability and systematic errors need to be considered conjointly in making a comparative evaluation of the conditions.

Table 6 (Part I) gives the frequency of statistically significant results by condition for Tables 3 and 4. Note the high frequency for C1 and appreciable drop in frequency for C2 and C3. The results for C3 are particularly impressive considered in conjunction with its high reliabilities.[3]

Part II, Table 6, summarizes the systematic error patterns, with P1 satisfying all models and P4 and P5 satisfying none. It is patterns P2 and P3 that show the value of the bias models, since these patterns indicate how C2 and C3—and the CB and RB models—fared for those cases in which C1, and hence the C model, was not satisfied.

P2 results show that of those cases showing systematic errors for C1 in Table 3 (7 cases) and Table 4 (9 cases), 3 of the 7 cases and 2 of the 9 cases do not show such errors for C2 *and* C3. Hence, for 5 of the 16 cases for which the C model failed by virtue of presence of systematic errors, the CB model was satisfied.

P2 and P3 results in conjunction show that of those cases showing systematic errors for C1 in Tables 3 and 4, 5 of the 7 cases (Table 3) and 6 of the 9 cases (Table 4) do not show systematic errors for C3. Hence, for 11 of the 16 cases for which the C model failed, the RB model was satisfied.

The comparative fit of the conditions and the reduction in systematic errors for C2 and C3 are demonstrated graphically for the key patterns, P2

Table 6
Systematic Errors

I. Frequency (Statistically Significant Results)			II. Patterns			
				Frequency		
Condition	Table 3	Table 4	Pattern	Table 3	Table 4	Models Satisfied
1	7/9	9/15	P1. C1-C2-C3	2	6	All
2	4/9	5/15	P2. $\overline{C1}$-C2-C3	3	2	CB and RB
3	2/9	3/15	P3. $\overline{C1}$-$\overline{C2}$-C3	2	4	RB
			P4. $\overline{C1}$-C2-$\overline{C3}$	0	2	None
			P5. $\overline{C1}$-$\overline{C2}$-$\overline{C3}$	2	1	None

Note—$\overline{Ci}$ denotes condition i not *satisfied, i.e., systematic errors statistically significant; Ci denotes condition i satisfied.*

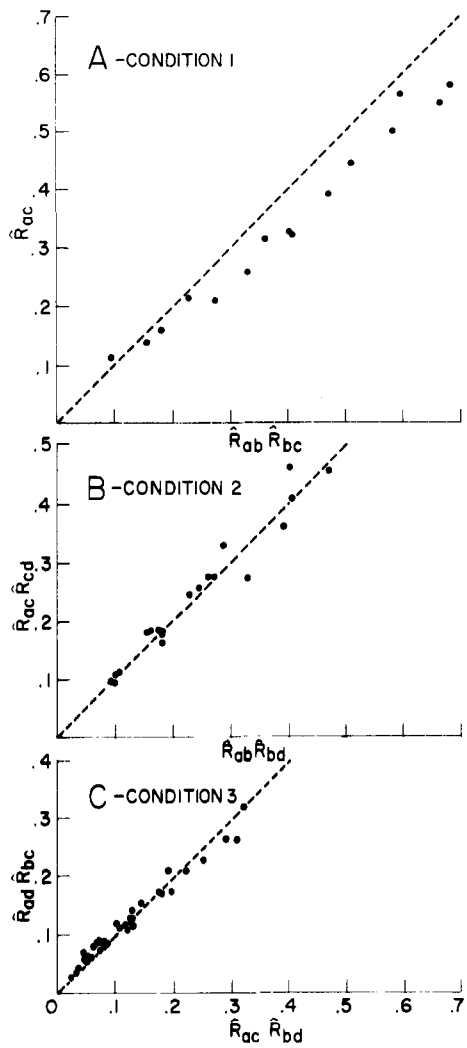

Figure 2. Comparative fit of conditions for Study No. 5 (Table 3), Pattern P2. Panel A, Condition 1: $\hat{R}_{ac}$ as a function of $\hat{R}_{ab}\hat{R}_{bc}$ (15 points). Panel B, Condition 2: $\hat{R}_{ac}\hat{R}_{cd}$ as a function of $\hat{R}_{ab}\hat{R}_{bd}$ (20 points). Panel C, Condition 3: $\hat{R}_{ad}\hat{R}_{bc}$ as a function of $\hat{R}_{ac}\hat{R}_{bd}$ (35 points).

and P3, with selected plots. Figure 2 gives plots of each condition for study No. 5, which shows systematic errors for C1 but not for C2 or C3—pattern P2. Panel A shows a plot of $\hat{R}_{ac}$ as a function of the product $\hat{R}_{ab}\hat{R}_{bc}$. If C1 is correct, the data points should fall randomly about the identity line. Obviously, the data points do not conform to C1, with all but one point falling well below the identity line. This plot is typical of the studies showing significant systematic errors for C1, although the *magnitudes* of the errors are much smaller than average.

Panel B of Figure 2 shows a plot of the product $\hat{R}_{ac}\hat{R}_{cd}$ as a function of the product $\hat{R}_{ab}\hat{R}_{bd}$; and Panel C shows a plot of the product $\hat{R}_{ad}\hat{R}_{bc}$ as a function of the product $\hat{R}_{ac}\hat{R}_{bd}$. Inspection of Panels B and C indicates reasonable conformity to C2 and C3, respectively, consistent with the results of the statistical tests.

Figure 3 illustrates pattern P3 with data from Study No. 3—showing significant systematic errors for C1 and C2 but not for C3. Figure 3 shows an excellent fit for C3 in the face of marked systematic errors for C1 and to a lesser degree, C2.

The dominant direction of systematic error for each condition is given by the following inequalities:

$$\text{C1: } R_{ab}R_{bc} > R_{ac} \tag{13}$$

$$\text{C2: } R_{ab}R_{bd} > R_{ac}R_{cd} \tag{14}$$

$$\text{C3: } R_{ac}R_{bd} > R_{ad}R_{bc};\ R_{ac}/R_{ad} > R_{bc}/R_{bd} \tag{15}$$

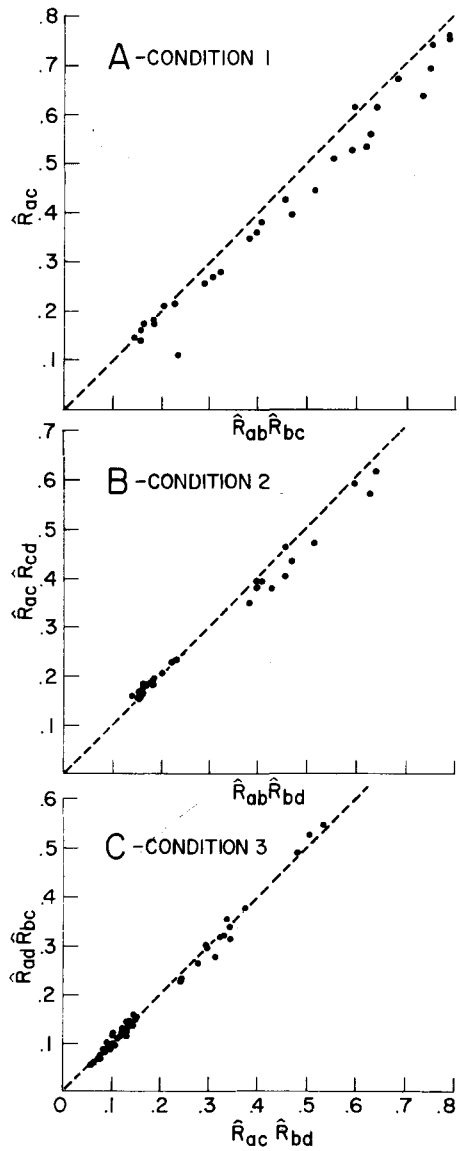

Figure 3. Comparative fit of conditions for Study No. 3 (Table 3), Pattern P3. Panel A, Condition 1: $\hat{R}_{ac}$ as a function of $\hat{R}_{ab}\hat{R}_{bc}$ (30 points). Panel B, Condition 2: $\hat{R}_{ac}\hat{R}_{cd}$ as a function of $\hat{R}_{ab}\hat{R}_{bd}$ (30 points). Panel C, Condition 3: $\hat{R}_{ad}\hat{R}_{bc}$ as a function of $\hat{R}_{ac}\hat{R}_{bd}$ [45 points with 30 points clustered below the point (.2, .2)]. The large number of points is due to combining results from three groups.

Table 7
Geometric Mean Estimates: β and β_b

Table 3 Study	$\hat{\beta}$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
1	1.029	1.058	1.019	1.000		
2	1.069	1.050	1.100	1.069	1.032	
3	1.248	1.394	1.343	1.124	.981	
4	1.150	1.107	1.074	1.225	1.248	1.172
5	1.369	1.466	1.391	1.433	1.291	1.064
6	1.374	1.512	1.469	1.309	1.241	1.159
7	1.387	1.323	1.466	1.294		
8	1.335	1.276	1.345	1.429		
9	.981	.897	1.021	1.312	.876	.620
Medians						
All Studies	1.248	1.276	1.343	1.294	1.137	1.112
Significant (C1)	1.335	1.323	1.345	1.309	1.241	1.159
Significant (C2)	1.309	1.430	1.367	1.267	1.137	1.112

These inequalities hold for all statistically significant cases, and hence the generalization can be made that if a structural condition is violated, the error is in one direction only, as given in Equations 13-15.

The estimation of the bias parameters and the relation of these parameters to error magnitudes will now be considered. The bias parameters β and β_b are defined by Equations 5 and 8, respectively. Each triple (a, b, c) provided an estimate of β, and the estimator, $\hat{\beta}$, was taken as the geometric mean of these estimates, i.e., the geometric mean of the $(R_{ab}R_{bc})/R_{ac}$. The estimators, $\hat{\beta}_b$ (b = 2, 3, $\cdots$, n − 1), were obtained by taking the geometric mean of all estimates $(R_{ab}R_{bc})/R_{ac}$ derived from triples (a, b, c) for which b is the middle stimulus.

Estimates of the bias parameters β and β_b are given in Table 7 for each study (Table 3) and in Table 8 for each subject (Table 4).

The magnitudes of β and β_b can be related to the direction of systematic errors, as follows: (1) If the CB or RB model holds but not the C model, then Inequality 13 is implied by $\beta, \beta_b > 1$ (holding in all 16 cases in which the test of C1 was significant). (2) If the RB model holds but not the CB model, then Inequality 14 is implied by a decreasing ordering on β_b, i.e., for β_b decreasing as b increases. Tables 7 and 8 indicate a trend toward a decreasing ordering, but the trend is far from perfect. Note that we cannot similarly explain the direction of systematic errors for C3, since the theory does not include a comparator model that may hold in the presence of violations of C3.

Bias parameters can also be used to examine the effect on scale values of using an incorrect model. (1) If the C model is used but the CB model is correct, the error in estimating Ψ_a/Ψ_b is $\beta - 1$. Inspection of Tables 7 and 8 indicates that the error can be quite large: For those cases for which systematic errors were significant for C1 but not for C2, the median error is 33.5%, extremely high. (2) Denote by $E_b(C1/C3)$ the error in estimating Ψ_a/Ψ_b if the C model is used but the RB model is correct. Then $E_b(C1/C3) = (\beta_b - 1)$ gives the error. Again, the errors are quite large. Defining the largest stimulus of each pair as a "standard," it can be stated that in general the tendency is for large errors for low standards and low errors for large standards. The CB model gives a kind of weighted mean of these errors. (3) The error made in estimating Ψ_a/Ψ_b using the CB model if it is wrong and the RB model is correct if $E_b(C2/C3) = (\beta_b/\beta) - 1$. In general, errors tend to be positive for low standards, small for middle standards, and negative for high standards. Positive errors mean that the scale value ratios are overestimated by the incorrect model, and negative errors that the ratios are underestimated. In the

Table 8
Geometric Mean Estimates: β and β_b

Table 4 Subjects	$\hat{\beta}$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
K.H.	1.205	1.810	1.068	.963	1.000
J.B.	1.022	.974	1.039	1.075	.961
D.S.	1.031	1.170	.967	.995	.948
D.K.	1.053	1.167	.995	1.066	.883
T.M.	1.053	1.467	.798	1.032	.950
R.H.	1.282	1.801	1.267	1.002	.999
T.T.	1.254	1.463	1.226	1.181	1.008
D.R.	1.264	1.301	1.429	1.166	1.023
D.S.R.	1.351	1.198	1.932	1.180	.989
D.L.	1.253	1.349	1.391	1.150	.949
L.C.	1.081	.794	1.214	1.340	1.007
W.H.	2.045	3.056	2.739	1.287	1.002
G.S.	1.400	1.329	1.850	1.261	.979
L.B.	1.390	2.037	1.329	1.095	1.036
M.W.	1.333	1.136	2.020	1.146	.972
Medians					
All Subjects	1.254	1.329	1.267	1.146	.989
Significant (C1)	1.333	1.349	1.391	1.166	.999
Significant (C2)	1.282	1.801	1.329	1.150	1.008

studies reviewed, incorrect application of the C model would, in general, result in large overestimation, whereas incorrect application of the CB model would result in smaller errors, both overestimation and underestimation.

The empirical comparison of models with differing numbers of parameters is not a trivial problem. In a typical paradigm, two models may be compared via a goodness of fit test based on the same set of data. If two models are equivalent except that the second involves an additional parameter, then of course the second *must* fit at least as well if the test is carried out on the same set of data and the parameters are used to improve the fit. Note, however, that for the present theory the parameters are not directly involved in the comparative tests, and the same set of data is not used to test each condition. The analyses are carried out directly on the structural conditions expressed entirely in terms of observables and are parameter-free. Therefore, it does *not* follow that in the reliability and systematic errors analysis the RB model *must* fare better than the CB and C models, and the CB better than the C model, by virtue of the additional parameters in the model equations. And, indeed, as shown in Tables 3 and 4, there are some reversals. Thus a significant advantage of the approach taken in this paper is the possibility of statistical comparisons of the models via structural conditions expressed entirely in terms of observables, rather than goodness-of-fit tests of model equations involving differing numbers of parameters.

Concluding Remarks

The present theory assumes that there exist "true" sensation ratios Ψ_a/Ψ_b for each pair (a, b). But if the C model is incorrect (invalidating the traditional method of ratio estimation), then the judgment R_{ab} is distorted by bias or context effects as measured by β or β_b. If $\beta,\beta_b > 1$, then $R_{ab} > \Psi_a/\Psi_b$, i.e., the judgment R_{ab} *overestimates* the sensation ratio Ψ_a/Ψ_b, an assimilation effect. If $\beta,\beta_b < 1$, then the judgment R_{ab} *underestimates* the sensation ratio, a contrast effect.

For all studies for which C1 systematic errors were significant (Table 3), and *all* subjects in Study No. 4 (Table 4), $\beta > 1$, implying an assimilation effect, i.e., the smaller stimulus was overestimated relative to the larger. For all studies for which C1 systematic errors were significant, $\beta_b > 1$ except $\beta_s < 1$ in Study No. 4. The same pattern holds for individual subjects in Study No. 4: $\beta_b > 1$ for significant subjects except $\beta_s < 1$ for some subjects. The trend toward a decreasing ordering on β_b leads to the interesting generalization that if the RB model is correct, then the assimilation effect is stronger for low standards (the larger member of the pair) and weaker for high standards.

Stevens and Galanter (1957) conjectured that ratio scales may be possible only on prothetic continua. Unfortunately, the studies analyzed include only one metathetic attribute—angles (Study No. 2)—for which C1 was rejected but C3 was satisfied. Eisler and Ekman (1959) constructed a ratio scale of pitch, a metathetic attribute, but used the method of fractionation, providing no data for a test of the structural conditions. We note from Table 3 that C1 was rejected for six of the eight prothetic continua, but that C3 was *satisfied* for six of the eight. Hence, these results fail to provide support for the Stevens and Galanter conjecture, but suggest, rather, that if the traditional method of ratio estimation (the C model) is used, then ratio scales are not possible even for prothetic continua.

Ratio scaling has received increasing theoretical scrutiny in recent years. A key question has been the kind of consistency ratio estimates must show in order to serve as the basis of measurement. This question has been approached in a number of indirect ways, e.g., via cross-modality matching, the psychophysical function (Marks, 1974), and inverse cross-modality matching (Lilienthal & Dawson, 1976).

The approach in this paper has been to ask what relations holding among the observable ratio judgments themselves are entailed by a scaling model, without recourse to another psychological attribute or a physical continuum. It was pointed out that precisely what relations must hold among ratio judgments depends on the assumed relationship between ratio judgments and psychological magnitudes. The postulation of three assumed relationships—models—led to three structural conditions (in addition to monotonicity) which provided a basis for choosing among the three models. Results showed that, in general, the C model—a characterization of the traditional ratio estimation scaling method—places constraints on the structure of ratio judgments (Condition 1, ratio consistency) that are too strong, and that the two weakened versions—the CB and RB models—need to be applied in most cases if ratio scales are to be constructed within the framework of the present theory. Condition 3, ratio constancy, shows generally high reliability and relative absence of systematic errors, justifying the utility of the RB model in the construction of ratio scales of sensation.

REFERENCE NOTE

1. Mashhour, M. *On the validity of scales derived by ratio and magnitude estimation methods.* Report of the Psychology Laboratory, University of Stockholm, 1961, No. 105.

REFERENCES

Bartko, J. J. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 1976, **83**, 762-765.

EISLER, H. Similarity in the continuum of heaviness with some methodological and theoretical considerations. *Scandinavian Journal of Psychology*, 1960, **1**, 69-81.

EISLER, H., & EKMAN, G. A mechanism of subjective similarity. *Acta Psychologica*, 1959, **16**, 1-10.

EKMAN, G. Two generalized ratio scaling methods. *Journal of Psychology*, 1958, **45**, 287-295.

EKMAN, G., GOUDE, G., & WAERN, Y. Subjective similarity in two perceptual continua. *Journal of Experimental Psychology*, 1961, **61**, 222-227.

ENGEN, T., & LINDSTRÖM, C. O. Psychophysical scales of the odor intensity of amyl acetate. *Scandinavian Journal of Psychology*, 1963, **4**, 23-28.

FAGOT, R. F., & STEWART, M. Tests of product and additive scaling axioms. *Perception & Psychophysics*, 1969, **5**, 117-123.

FAGOT, R. F., STEWART, M. R., & KLEINKNECHT, R. E. Representations for biased numerical judgments. *Perception & Psychophysics*, 1975, **17**, 309-319.

GOUDE, G. *On fundamental measurement in psychology*. Stockholm: Almqvist & Wiksell, 1962.

KRANTZ, D. H. A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology*, 1972, **9**, 168-199.

LILIENTHAL, M. G., & DAWSON, W. E. Inverse cross-modality matching: A test of ratio judgment consistency for group and individual data. *Perception & Psychophysics*, 1976, **19**, 252-260.

MARKS, L. E. *Sensory processes, the new psychophysics*. New York: Academic Press, 1974.

MASHHOUR, M. *Psychophysical relations in the perception of velocity*. Stockholm: Almqvist & Wiksell, 1964.

MULAIK, S. A. *The foundations of factor analysis*. New York: McGraw-Hill, 1972.

SJÖBERG, L. Three models for the analysis of subjective ratios. *Scandinavian Journal of Psychology*, 1971, **12**, 217-240.

STEVENS, S. S. Problems and methods of psychophysics. *Psychological Bulletin*, 1958, **55**, 177-196.

STEVENS, S. S., & GALANTER, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, **54**, 377-411.

NOTES

1. Each R_{ab} was the average of several observations. Raw data for the studies in Table 3 were not available to the author.

2. For C1 and C2, special problems arose in the selection of stimulus sets. Consider the triples for C1: For each pair (a, c), the number of sets depends on the number of stimuli, b, between a and c (which, in the studies reported on, varied from one to five). In order to give equal weight to each pair (a, c), exactly one stimulus b was selected at random from those stimuli between each a and c. The problem for C2 was that if all tetrads were used for testing, then some $Y = R_{ab}R_{bd}$ would be used more than once. Hence, a method of selecting the stimulus set for C2 was devised that produced the largest number of tetrads satisfying the condition that each Y could be used only once. No such problems arose with C3, and hence all tetrads were included in the stimulus sets.

3. Fagot and Stewart (1969) also reported a poor fit to C1 for brightness judgments. However, the method of selecting independent triples for testing C1 did not produce stimulus sets that could have been used to test C2 and C3. Since the three conditions could not be compared, the study was omitted.

(Received for publication October 21, 1977;
revision accepted June 16, 1978.)