Journal of Experimental Psychology
1972, Vol. 93, No. 1, 35–42

MORALITY JUDGMENTS:

TESTS OF AN AVERAGING MODEL [1]

MICHAEL H. BIRNBAUM [2]

University of California, Los Angeles

Pairs of items describing objectionable behaviors were rated for their overall morality. Contrary to additive or constant-weight averaging models, the ratings were demonstrated to depend upon the range as well as the average scale value of the component behaviors. A range model accounted for more than half of the variance left unexplained by the additive models. One interpretation of the range effect postulates that each component stimulus produces a distribution of values. The value of the stimulus combination is assumed to be the mean value in the overlap of the component distributions, which is closer to the item with the narrower dispersion.

How immoral is it to both "pocket the tip the previous customer left for the waitress" *and* "poison your neighbor's dog whose barking bothers you?" Anderson (1968b) has suggested that *S*s combine information by averaging psychological values associated with each component stimulus. Thus, *S* independently assesses the morality of pocketing the tip and the morality of poisoning the dog and then averages his two assessments to arrive at an overall rating. According to this view, the psychological value of the *whole* is simply an average of the psychological values of the *parts*. The theory is a special case of a general additive model (Rosenberg, 1968), which can be written:

$$\Psi_{1\ldots k\ldots n} = \sum_{k=0}^{n} w_k s_k + c \qquad [1]$$

where $\Psi_{1\ldots k\ldots n}$ is the psychological value of the combination of n stimuli, s_k is the psychological value of stimulus k, s_o and w_o are the scale value and weight of the initial impression, and c is an additive constant. In the usual application of such models, the weights, w_k, are assumed to be independent of the scale values, and the effect of each stimulus is assumed to be independent of the other stimuli with which it is presented (Anderson, 1970). The constant-weight averaging model (Anderson, 1968b; Rosenberg, 1968) is a special case of the additive model and requires the additional assumption that the weights sum to unity.

The overt response, $R_{1\ldots k\ldots n}$, is assumed to be a monotonic transformation, J, of the psychological impression values:

$$R_{1\ldots k\ldots n} = J(\Psi_{1\ldots k\ldots n}). \qquad [2]$$

Since J is usually assumed to be linear, analysis of variance may be applied directly to the responses obtained in a factorial experiment to test the general additive model (Anderson, 1968b, 1970, 1971). These models predict no interaction between component stimuli, and the test for interactions provides a test of all of the special cases of this model including both additive and constant-weight averaging models.

These models have provided a good fit to the data obtained in previous studies using verbal stimuli (Anderson, 1968b, 1971; Slovic & Lichtenstein, 1971) and have received some support in studies using psychophysical stimuli (Anderson, 1970; Slovic & Lichtenstein, 1971). However,

[1] The author wishes to express his gratitude to Allen Parducci and Norman H. Anderson for comments and suggestions on earlier drafts of this paper. Support was received from National Institutes of Health Grant HD 00923; computing assistance was received from Campus Computing Network, University of California, Los Angeles. Thanks are due to Leslie Freedman and Clairice Veit for their assistance with Exp. I and to Andrew Comrey for providing the successive intervals scaling computer program.

[2] Requests for reprints should be sent to Michael H. Birnbaum, Department of Psychology, University of California, Los Angeles, California 90024.

TABLE 1

Morality Values for the 16 Component Items

Item no.	Item	Mean judgment	*SD*	Thurstone value	Thurstone dispersion
1	Keeping a dime you find in a telephone booth.[b,d]	8.79	.68	5.50	1.81
2	Playing poker on Sunday.[a,c]	8.78	.84	5.37	1.79
3	Cheating at solitaire.[c]	7.89	1.58	3.66	1.54
4	Wearing shorts on the street where it is illegal.[d]	7.81	1.43	3.38	1.50
5	Stealing a loaf of bread from a store when you are starving.[d]	7.75	1.73	3.52	1.52
6	Registering in a hotel under a false name.[c]	7.70	1.83	3.49	1.51
7	Stealing towels from a hotel.[b,c]	7.25	1.65	2.91	1.43
8	Failing to pay your bus fare when the conductor overlooks you.[a,c]	7.04	1.74	2.78	1.41
9	Failing to put back in the water lobsters which are shorter than the legal limit.[d]	6.54	2.03	2.44	1.34
10	Getting your own way by playing on other people's sympathies.[b,d]	5.78	2.05	2.08	1.31
11	Contributing money to a cause in which you do not believe in order to escape criticism.[a,d]	5.33	2.27	1.87	1.28
12	Pocketing the tip the previous customer left for the waitress.[c]	4.85	2.04	1.72	1.25
13	Bawling out servants publicly.[b,d]	4.58	1.96	1.50	1.22
14	Habitually borrowing small sums of money from friends and failing to return them.[a,c]	4.32	1.77	1.40	1.21
15	Spreading rumors that an acquaintance is a sexual pervert.[b,c]	3.27	.98	.85	1.13
16	Poisoning a neighbor's dog whose barking bothers you.[a,d]	1.95	1.46	.00	1.00

[a] First item of Exp. I.
[b] Second item of Exp. I.
[c] First item of Exp. II.
[d] Second item of Exp. II.

exceptions to the model have been obtained with psychophysical stimuli (Birnbaum, Parducci, & Gifford, 1971; Parducci, Thaler, & Anderson, 1968). Birnbaum et al. (1971) proposed a range model to account for interactions obtained in these experiments.

The range model asserts that the psychological impression of the stimulus combination is directly related to the *range* of the stimuli as well as the *mean*. For two stimuli, *i* and *j*, the model can be written:

$$\Psi_{ij} = ws_i + (1 - w)s_j + \omega|s_i - s_j|/(s_i + s_j), \quad [3]$$

where Ψ_{ij} is the impression of the pair, w and $(1 - w)$ are the mathematical weights associated with order of presentation, s_i and s_j are the scale values of the stimuli, and ω is an empirical constant which represents the magnitude of the range effect. The scale values are assumed to all have the same sign.

The present experiments provide a test of the applicability of the general additive model to morality judgments and permit a comparison with the range model.

Method

In both experiments, *S*s were instructed to read each pair of behavior items and judge "how wrong that *pair* of actions would be in terms of your own personal set of values."

Behavior items.—Brief characterizations of 16 immoral behaviors were selected from those used by Parducci (1968). These single actions were judged by 81 undergraduates who made their ratings on a 9-point scale. The category labels were the same as in Exp. II below. The mean category judgments are presented as the scale values in Table 1, along with the corresponding Thurstone Case 6 values (Bock & Jones, 1968).

Subjects.—The *S*s were 219 University of California, Los Angeles, undergraduates; 81 *S*s judged the single items for the preliminary scaling. There were 100 *S*s in Exp. I and 38 different *S*s in Exp. II. The *S*s were run in groups of from 2 to 10 *S*s each.

Experiment I.—The stimuli consisted of 25 pairs of items produced from a 5 × 5 factorial design in which the first and second items could each take five values. Table 1 lists these items. The 25 pairs were printed in booklets in a single random order, and *S*s were instructed to read the entire list before making their judgments.

The ratings were recorded in numerical form, using a 9-point scale anchored by five category labels: 1—"neither good nor bad," 3—"questionable, but not particularly bad or wrong," 5—"undesirable, a good person would not do this," 7—"wrong, highly questionable," 9—"seriously wrong." To facilitate comparison with related work, judgments obtained in both experiments were subtracted from 10 to reverse the scale.

Experiment II.—The two experiments differed in the labeling of the response scale and also in the particular stimulus pairs. The labels were adjusted for Exp. II to permit more categories at the "seriously wrong" end of the scale. The new labels were: 1—"not particularly bad or wrong," 3—"undesirable, a good person would not do this," 5—"wrong, highly questionable," 7—"seriously wrong," 9—"extremely evil." The stimuli of Exp. II consisted of 64 pairs of items, produced from an 8 × 8 factorial design using all 16 items listed in Table 1. Each *S* received a booklet in which the pairs were printed in one of four random orders. The *S*s judged each pair twice; the order of the pairs and the order of items within each pair was reversed for the second replicate. The order of items within each pair on the first replicate was counterbalanced on half of the forms.

Results and Discussion

Figure 1 plots the mean rating obtained in both experiments against the best-fit values of the averaging model. The model is fitted separately for each experiment by a least-squares solution for the weights and additive constant $(c + w_o s_o)$ of Equation 1, using the prior scale values of Table 1. The multiple correlations for Exp. I and II were .938 and .952, respectively. These high coefficients indicate that the additive (or averaging) model provides a satisfactory basis for practical prediction, and they also reflect high reliability of the scaling. However, any model predicting that response varies directly with each of the component values would yield high correlations (Anderson, 1968b; 1971). Therefore, it is necessary to examine the deviations from the model.

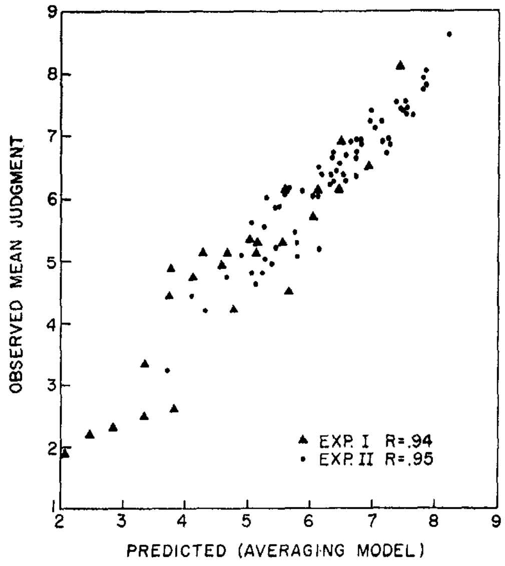

FIG. 1. Observed mean judgments as a function of averaging model predictions.

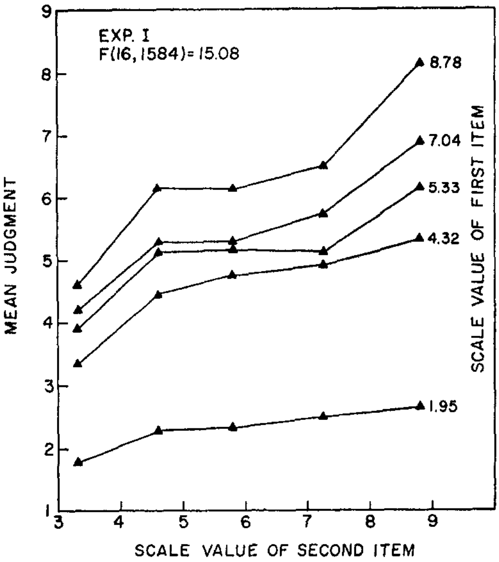

FIG. 2. Mean judgment of each pair of behaviors of Exp. I as a function of the scale values of the first and second items.

Figure 2 shows the mean judgments of the 25 pairs of Exp. I. Each curve represents the mean judgments of five pairs with the same first item. The slope of each curve represents the effects of the second item; the distances between the curves represent the effects of the first item. According to additive (or averaging) models, the curves should be parallel. Instead, they diverge to the right. This divergence is characteristic of the individual data for 85 of the 100 *S*s. The analysis of variance test of the interaction is highly significant, F (16, 1584) = 15.08, $p < .001$. This interaction is inconsistent with the

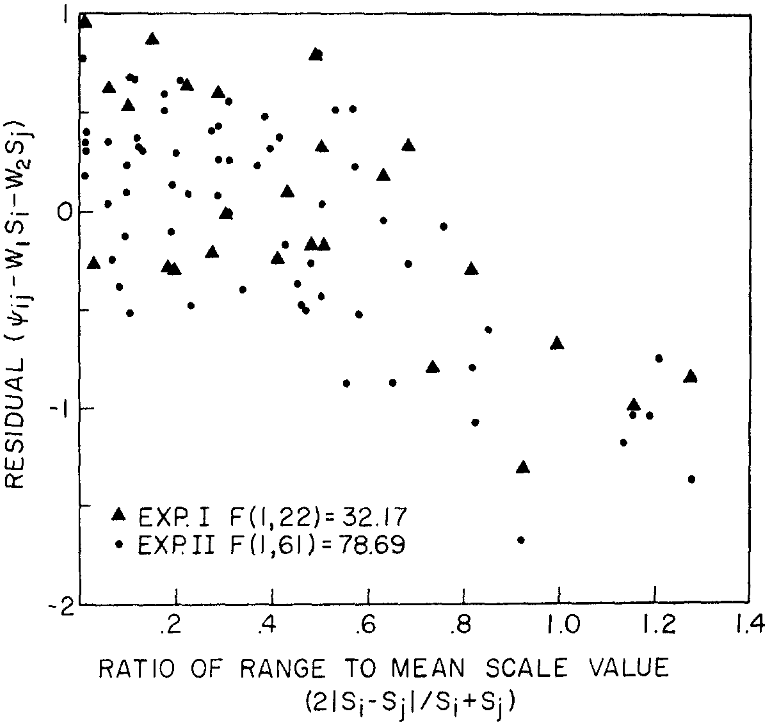

FIG. 3. Residuals from the best fit of the averaging model (following transformation) as a function of the ratio of the within-set range to the mean scale value of the pair.

general additive model (Equation 1) and therefore provides evidence against all of the special cases of this model (Rosenberg, 1968).

One possibility is that these interactions (deviations from additivity) reflect compression of the response scale at the "seriously wrong" end, i.e., nonlinearity in the function J of Equation 2. As a check on the possibility of this "end" effect, all pairs containing either of the two items with the lowest scale values were removed from the analysis, which was thus reduced to a 4 × 4 design. The interaction remained highly significant, F (9, 891) = 8.35, $p < .001$.

The data of Exp. II provided a second check on the response scale. Although the change in scale labels for Exp. II reduced the use of the lower categories (as shown in Fig. 1), the interaction had the same form as that of Exp. I and was highly significant, F (49, 1813) = 4.21, $p < .001$.

Monotonic Transformation

As a final check on the response scale, the responses were transformed to impression values (Ψ in Equation 2) by the procedure described in Birnbaum et al. (1971). This procedure assumes the validity of the prior scale values and the validity of the averaging model for pairs of minimal within-set range. The transformation, which relates the judgments of pairs to impression values on the prior scale, is then estimated as a polynomial by a least-squares criterion. Although this procedure permits radical rescaling, the best-fit transformation was fairly linear so that the interactions retained the same form and remained highly significant. The finding that the ratings of pairs of minimal within-set range are linearly related to the ratings of the single items suggests that the failure of the additive and averaging models is not due to improper scaling of the responses.

Range Model

Figure 3 plots the residuals from the additive model of Equation 1 (following transformation) as a function of the ratio of the range to the mean scale value for each pair. Since these residuals are simply "obtained minus predicted," they should be zero in theory and at least show no trend. Instead, the relationship follows the linear trend predicted by the range model (Equation 3). In addition, the transformed data from both experiments yield approximately the same estimate of ω for Exp. I and II (-2.61 and -2.32); i.e., both sets of points fall approximately on the same line. The partial correlations with the range term, $-.77$ and $-.75$ for the two sets of data are highly significant, $F(1, 22) = 32.17$ and $F(1, 61) = 78.69$. The sum of squared errors for the additive or constant-weight averaging models is more than twice the sum of squared errors for the range model.

Theoretical Interpretations

Overlap of values.—A possible theoretical basis for the range model assumes that each component stimulus produces a distribution of values on an underlying psychological continuum. This hypothetical distribution represents all the possible values of the component stimulus. The present hypothesis assumes that S bases his rating of a single component on the mean of its distribution of possible values; when instructed to judge a pair, he estimates the mean of those values common to both distributions, i.e., their overlap. If the dispersion of each component's distribution of possible values varies directly with its mean, the value of the combination will be related to the range (i.e., difference between the means of the components) as well as to the average of the two mean values. As shown in Fig. 4, the rating of the combination will be shifted toward the component with narrower dispersion; the relative magnitude of the shift increases with the difference in dispersion. For the present data, this hypothesis requires that the immoral items have narrower dispersions than the milder items.

Application of successive interval scaling (Torgerson, 1958, pp. 205–210) to the ratings of the 16 single items revealed that scale value and dispersion were indeed correlated ($r = .82$). This correlation suggested the use of Thurstone Case 6 values in Fig. 4 to illustrate the hypothesis. Since the function relating Thurstone values to mean judgments is negatively accelerated, the effect of the difference in dispersion is inversely proportional to the mean scale value.

Range effects have been reported in previous studies of information integration. Willis (1960) found that judgments of the attractiveness of groups of facial photographs varied inversely with the "heterogeneity" of the component photograph scale values and directly with their mean. Heterogeneity was defined as the average absolute deviation of the components around their mean and is therefore analogous to the range term of Equation 3. Weiss (1963) found similar interactions for judgments of opinion triplets. Judgments of intelligence show interactions due to "cue inconsistency," a concept which is also analogous to that of range (Slovic & Lichtenstein, 1971).

The range model has also been applied to judgments of the likableness of hypothetical persons described by pairs of adjectives, including adjectives from both ends of the likableness continuum. The range term accounted for 80% of the variance left unexplained by the additive models (Birnbaum, 1970, 1971). Work is currently under way to investigate the relative effects of moral and immoral items on integrated impressions of morality.

Other interpretations.—Another interpretation of the interactions postulates shifts in the components themselves. When separate ratings of the components presented in the set (Anderson, 1968b, 1971; Birnbaum et al., 1971) shift systematically, it is usually toward the values of the other components of the same set. However, if the magnitude of shifts were simply a linear function of the values of the com-

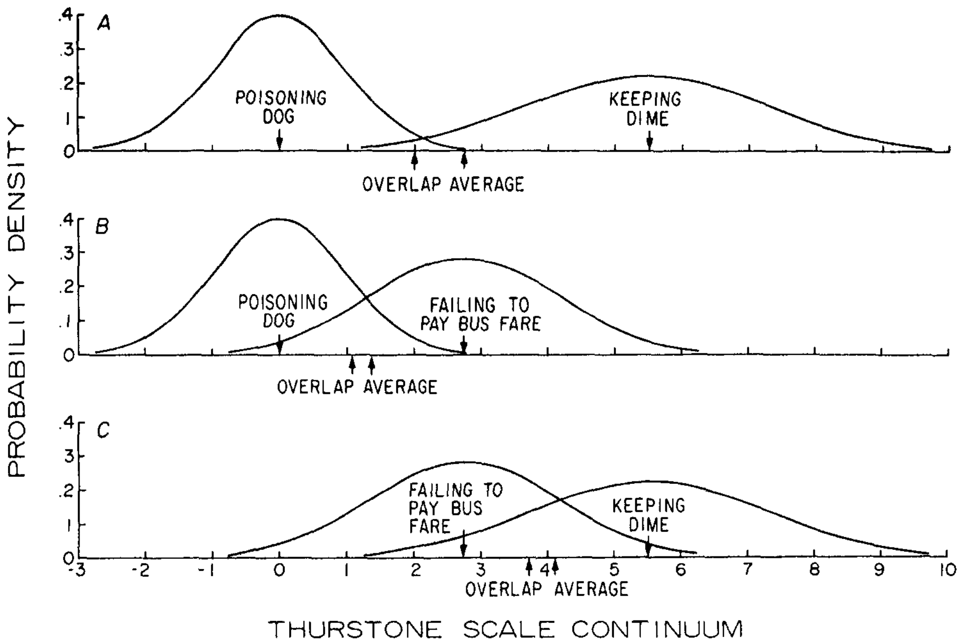

FIG. 4. Overlap of value hypothesis illustrated. (Arrows labeled "average" represent the averaging model predictions; each arrow labeled "overlap" represents the mean of the overlap of two distributions.)

ponents, this contextual effect would not produce interactions (Anderson, 1968b, 1971).

Instead of the items changing value in context, it is possible that the weight or importance of each component depends upon its value (Anderson, 1971) and perhaps also on the values of the other components of the same set. Osgood and Tannenbaum (1955) have suggested a model in which the weight of an item is proportional to its extremity. Differential weighting has also been considered by Manis, Gleason, and Dawes (1966) and by Feldman (1968). There are many possible explanations for differential weighting models. For the present data, the more serious misbehaviors might receive greater relative weight in determining the overall impression because they require action such as punishment of the wrongdoer. Similarly, if serious misbehaviors occur less often, these items would provide more information and consequently receive greater weight. Weighting could also depend upon the reliability of the information, with weight inversely proportional to dispersion (as in Fig. 4).[3]

Differential weighting models might be consistent with the present interactions which contradict the constant-weight models. These more complicated averaging models require the estimation of two parameters for each stimulus, weight and scale value, whereas the constant-weight models (with appropriate experimental designs) require the explicit estimation of only the scale values. The number of parameters can be reduced by assuming

[3] The distributional interpretation shown in Fig. 4 can provide the basis for the derivation of an averaging model with differential weights. Let s_i and σ_i be the mean and standard deviation for Stimulus i. If it is assumed that the combined impression, Ψ, of k stimuli is the value that maximizes the product of the normal probability densities (as in Fig. 4), then it can be shown that $\Psi = \sum_{i=1}^{k} (s_i/\sigma_i) / \sum_{i=1}^{k} (1/\sigma_i)$, which is an averaging model with weights inversely proportional to dispersion. For two stimuli, this criterion also implies that $P(M > \Psi | \text{Stimulus } 1) = P(M < \Psi | \text{Stimulus } 2)$, where M is morality.

that weight is a linear function of scale value. This simplified variable-weight model does as well as the range model for the data of Exp. I, but poorer for Exp. II. After fitting this variable-weight model to the data of Exp. II, the range term accounts for a significant portion of the unexplained variance, F (1, 60) = 22.6, $p < .001$.

The distinction between differential weighting models and the constant-weight additive or averaging models cannot be emphasized too strongly. The additive models (Rosenberg, 1968) require many fewer parameters but cannot account for interactions. The differential weighting models can account for interactions (Anderson, 1971) but suggest a different process of information integration and raise the theoretical problem of explaining the nature and cause of the weighting.

On the other hand, it would be difficult at present to make a sharp distinction between the range model and the differential weighting models on the basis of existing data, since both models can predict simple convergent or divergent interactions. For example, recent experiments by Oden and Anderson (1971) suggest that judgments of the "badness" of groups of criminals are inconsistent with additive models of information integration. A differential weighting model was applied to the data, although the deviations from additivity were in the direction that would be predicted by the range model.

Conclusions

The present data suggests that *S*s do not combine information about the morality of behaviors by simply summing or averaging the values associated with the components. The interactions obtained in both experiments are consistent with other evidence (e.g., Anderson, 1965, 1968a; Feldman, 1968; Lampel & Anderson, 1968; Oden and Anderson, 1971) suggesting that integrated impressions are shifted toward the evaluatively lower information contained in the set.

The present experiments do not preclude the possibility that the interactions are due to contextual changes in value or weight. Differential weighting of the stimuli may account for the shifts but the range model appears to provide a better fit to the data.

The range model can be interpreted as reflecting an underlying psychological representation of the stimulus items in which the lower valued stimuli are located more precisely than the neutral ones, so that the stimulus combination falls closer to the lower items.

References

Anderson, N. H. Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, 1965, **70**, 394–400.

Anderson, N. H. Application of a linear-serial model to a personality-impression task using serial presentation. *Journal of Personality and Social Psychology*, 1968, **10**, 352–362. (a)

Anderson, N. H. A simple model for information integration. In R. P. Abelson et al. (Eds.), *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally, 1968. (b)

Anderson, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970, **77**, 153–170.

Anderson, N. H. Information integration and attitude change. *Psychological Review*, 1971, **78**, 171–206.

Birnbaum, M. H. Impression formation: Averaging versus range model. Paper presented at the Mathematical Psychology Meeting, Miami, September 1970.

Birnbaum, M. H. Impression formation: Difference judgments as a basis for response rescaling. Paper presented at the meeting of The Western Psychological Association, San Francisco, April 1971.

Birnbaum, M. H., Parducci, A., & Gifford, R. K. Contextual effects in information integration. *Journal of Experimental Psychology*, 1971, **88**, 158–170.

Bock, R. D., & Jones, L. V. *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day, 1968.

Feldman, S. What do you think of a *cruel, wise* man? The integrative response to a stimulus manifold. In R. P. Abelson et al. (Eds.), *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally, 1968.

Lampel, A. K., & Anderson, N. H. Combining visual and verbal information in an impression formation task. *Journal of Personality and Social Psychology*, 1968, **9**, 1–6.

Manis, M., Gleason, T. C., & Dawes, R. M. The evaluation of complex social stimuli. *Journal of Personality and Social Psychology*, 1966, **3**, 404–419.

Oden, G. C., & Anderson, N. H. Differential weighting in integration theory. *Journal of Experimental Psychology*, 1971, **89**, 152–161.

Osgood, C. E., & Tannenbaum, P. H. The principle of congruity in the prediction of attitude change. *Psychological Review*, 1955, **62**, 42–55.

Parducci, A. The relativism of absolute judgments. *Scientific American*, 1968, **216**, 84–90.

Parducci, A., Thaler, H., & Anderson, N. H. Stimulus averaging and the context for judgment. *Perception & Psychophysics*, 1968, **3**, 145–160.

Rosenberg, S. Mathematical models of social behavior. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology*. Vol. 1. (2nd ed.) Reading, Mass: Addison-Wesley, 1968.

Slovic, P., & Lichtenstein, S. C. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, **6**, 649–744.

Torgerson, W. *Theory and methods of scaling*. New York: Wiley, 1958.

Weiss, W. Scale judgments of triplets of opinion statements. *Journal of Abnormal and Social Psychology*, 1963, **66**, 471–479.

Willis, R. H. Stimulus pooling and social perception. *Journal of Abnormal and Social Psychology*, 1960, **60**, 365–373.

(Received June 18, 1971)