Journal of Experimental Psychology
1973, Vol. 99, No. 3, 395–399

MORALITY JUDGMENT: TEST OF AN AVERAGING MODEL WITH DIFFERENTIAL WEIGHTS[1]

MICHAEL H. BIRNBAUM[2]

University of California, Los Angeles

Ratings of persons described by sets of moral and immoral actions were inconsistent with additive and averaging models of information integration. An averaging model with differential weights could not give a consistent account of the effects of both the number of items and the heterogeneity of the items in the set. Highly immoral deeds appear to have an overriding influence on the overall judgment: Having committed one bad deed, a person will be rated "bad," with his good deeds having little influence. Morality judgment may thus represent a truly configural process.

Recent research with judgments of the morality of objectionable behaviors (Birnbaum, 1972a) suggests that *S*s integrate evaluations of immorality in a nonadditive fashion. Contrary to additive or constant-weight averaging models, judgments of the overall morality of 2 actions depend upon the range of the values within the set, as well as their sum or mean. The greater the range of the items, holding mean scale value constant, the lower the judgment of morality.

The interactions obtained by Birnbaum (1972a) were interpreted as consistent with either a range model (Birnbaum, Parducci, & Gifford, 1971) or an averaging model with differential weights (Anderson, 1972; Oden & Anderson, 1971). The differential-weight averaging model could account for the interactions with the assumption that the more immoral items have greater weight.

The present study extends the investigation of morality judgment to test the differential-weight averaging model. This is accomplished by using items of both moral and immoral value and by varying the number of items in the set.

METHOD

The *S*s were required to make a rating of the overall morality of performing *all* of the behaviors described in each set of items. They recorded their ratings, 1 through 9, to represent different verbal categories: 1 = very very bad, 2 = very bad, 3 = bad, 4 = slightly bad, 5 = neutral (neither good nor bad), 6 = slightly good, 7 = good, 8 = very good, and 9 = very very good.

Subjects. The *S*s were 60 undergraduates at the University of California, Los Angeles, fulfilling a requirement in introductory psychology.

Stimuli. The 16 items used to construct the sets are listed in Table 1. The mean judgments and standard deviations listed in the table were obtained in preliminary work[3] in which 101 *S*s rated the morality of each of 300 single items using the same 9 categories. Table 1 shows the 4 items chosen for each of 4 levels of morality: low, moderately negative, moderately positive, and high, indexed by L, M^-, M^+, and H, respectively. The replicates within each level are labeled A, B, C, and D.

Design. In addition to the 16 single items, there were 50 sets of items produced from combinations of the items in Table 1. The combinations are

[1] Thanks are due to Allen Parducci, Norman H. Anderson, and Dwight Riskey for their helpful comments on earlier versions of this paper. Special thanks to Clairice T. Veit for her assistance with the experimental work and her suggestions on the manuscript. This research was completed while the author held a National Defense Education Act Title IV graduate fellowship at University of California, Los Angeles; this paper was completed while the author held a National Institute of Mental Health postdoctoral fellowship at University of California, San Diego. Computing assistance was received from Campus Computing Network, University of California, Los Angeles. Assistance in the preparation of the manuscript was provided by Center for Human Information Processing through National Institute of Mental Health Grant MH-15828.

[2] Requests for reprints should be sent to Michael H. Birnbaum, Department of Psychology, University of California, San Diego, P.O. Box 109, La Jolla, California 92037.

[3] The author gratefully acknowledges the assistance of Clairice T. Veit, Herbert Marsh, and Howard Fleiner, who wrote many of the stimulus items, conducted the preliminary scaling experiments, and kindly provided the normative data for the items. Many of the items of immoral value were taken from McGarvey (1943).

TABLE 1

MORALITY ITEMS AND PRIOR SCALE VALUES

Item label[a]	Item	$\bar{X}$	SD
A(L)	Putting razor blades in children's apples on halloween	1.54	1.57
B(L)	Torturing prisoners of war to extract information	2.07	1.48
C(L)	Selling pork from trichinotic pigs	1.68	1.28
D(L)	Using guns on striking workers	1.71	.91
A(M^-)	Faking your data in a scientific experiment	3.41	1.21
B(M^-)	Bribing your way out of a summons for speeding	3.55	1.32
C(M^-)	Turing in a false fire alarm	3.01	1.88
D(M^-)	Writing anonymous letters to frighten a personal enemy	3.03	1.16
A(M^+)	Helping an old lady to cross the street	6.77	1.19
B(M^+)	Visiting a sick friend in the hospital	6.95	.99
C(M^+)	Fixing your friend's car for free	6.62	1.07
D(M^+)	Planting flowers to beautify the neighborhood	6.57	1.05
A(H)	Donating a kidney to a child needing an organ transplant	8.13	1.16
B(H)	Rescuing a family from a burning house	8.40	.75
C(H)	Talking a friend out of suicide	8.14	.96
D(H)	Preventing a forcible rape	8.15	1.34

[a] Levels of morality: L = low; M^- = moderately negative; M^+ = moderately positive; and H = high. The replicates in each level are labeled A, B, C, and D.

listed in Table 2, which segregates the sets into homogeneous and heterogeneous sets of each set size. Each set of letters refers to a set of items; for example D(M^-)B(H) would refer to the set of items, "writing anonymous letters to frighten a personal enemy *and* rescuing a family from a burning house." The order in which the items were printed within each set is indicated by the order of the letters in Table 2.

The combinations were chosen to provide factorial designs of 2 types: (*a*) Set Size × Scale Value, in which the sets were composed of items of homogeneous scale value, but with either 1, 2, 3, or 4 items; and (*b*) Scale Value × Scale Value, in which each set of 2, 3, or 4 items contained items of varying levels of morality.

Procedure. The 66 sets of items were printed in random order on 4 pages, with between 15 and 18 sets per page. The pages were ordered in all possible permutations to produce booklets.

The cover page of each booklet contained the written instructions and response scale. The instructions read (in part) as follows:

> . . . Your task is to read each set of actions and then judge how "good" or "bad" it would be to carry out *all* of the actions. . . . In other words, . . . how morally "commendable" or "reprehensible" a person would be who carried out *all* of the actions. . . . Be sure to read *all* of the actions and consider them of equal importance in forming your overall impressions of morality. . . . Read through the first page or two to get an idea of the nature of the items before you begin to record your judgments.

RESULTS

Figure 1A shows the mean judgments of pairs of items constructed from the A × B factorial design (i.e., the judgments of the 16 sets of 2 items formed from all combinations of replicate A and B items, including homogeneous and heterogeneous sets). The slopes of the curves represent the effects of the B item; the vertical distances between the curves represent the effects of the A item. According to additive or constant-weight averaging models, the curves should be parallel (Anderson, 1968). Instead, the curves diverge to the right. This divergence is characteristic of the data for 49 of the 60 *S*s, and the analysis of variance test for the interaction is highly significant, F (9, 531) = 38.28.

The sets of 3 and 4 items were constructed to provide 6 additional 2 × 2 factorial designs to permit additional

TABLE 2

SETS USED IN EXPERIMENT

Set size	Set composition	
	Homogeneous	Heterogeneous
2	A(i)B(i); i = L,M^-,M^+,H C(i)D(i); i = L,M^-,M^+,H	A(i)B(j); $i \neq j$; i = L,M^-,M^+,H; j = L,M^-,M^+,H C(L)D(H), C(H)D(L)
3	A(i)B(i)C(i); i = L,M^-,M^+,H C(i)D(i)A(i); i = L,M^-,M^+,H	A(L)B(H)C(H), A(H)B(L)C(L) C(L)D(H)A(H), C(H)D(L)A(L)
4	A(i)B(i)C(i)D(i); i = L,M^-,M^+,H C(i)D(i)A(i)B(i); i = L,M^-,M^+,H	A(L)B(L)C(H)D(H), A(H)B(H)C(L)D(L) A(L)B(H)C(H)D(H), A(H)B(L)C(L)D(L) C(L)D(L)A(H)B(H), C(H)D(H)A(L)B(L) C(L)D(H)A(H)B(H), C(H)D(L)A(L)B(L)

Note. Each set of letters refers to a set of items, and the order in which the items were printed within each set is indicated by the order of the letters. Levels of morality: L = low; M^- = moderately negative; M^+ = moderately positive; and H = high. The replicates within each level are labeled A, B, C, and D.

evaluation of the additive models. All 6 interactions were of the same form as Figure 1A and all were highly significant. The interactions are very large and all of the same divergent form as those obtained previously with sets composed only of immoral items (Birnbaum, 1972a). These results thus provide further evidence against the additive and constant-weight averaging models for morality judgment. They show that the divergent interaction is found for sets of 2, 3, and 4 items, and for sets that include items of both moral and immoral value.

The simple divergent interaction is also inconsistent with the special cases of averaging models with differential weighting proposed by Osgood and Tannenbaum (1955) and Manis, Gleason, and Dawes (1966). These models assume that weight is related to the extremity of the information and therefore predict that the curves in Figure 1A should converge equally at both ends.

Figure 1B plots mean judgments of the sets of items of homogeneous scale value as a function of level of scale value (spaced according to the marginal means) with a separate curve for sets of each set size. The fact that response becomes more extreme as the number of items is increased is consistent with previous findings (Anderson, 1965, 1967; Fishbein & Hunter, 1964).

However, these data are inconsistent with the additive model (Fishbein & Hunter, 1964), the constant-weight averaging model (Anderson, 1967), and also the model of Manis et al. (1966), all of which imply that the curves of Figure 1B should form a set of straight lines, intersecting at a common point. Instead, there is a pinching at the endpoints, and the residual from the bilinear interaction is statistically significant, F (8, 472) = 9.34. Although this residual interaction is inconsistent with the idea that the items have equal weight, it is consistent with greater weighting for the more extreme items.

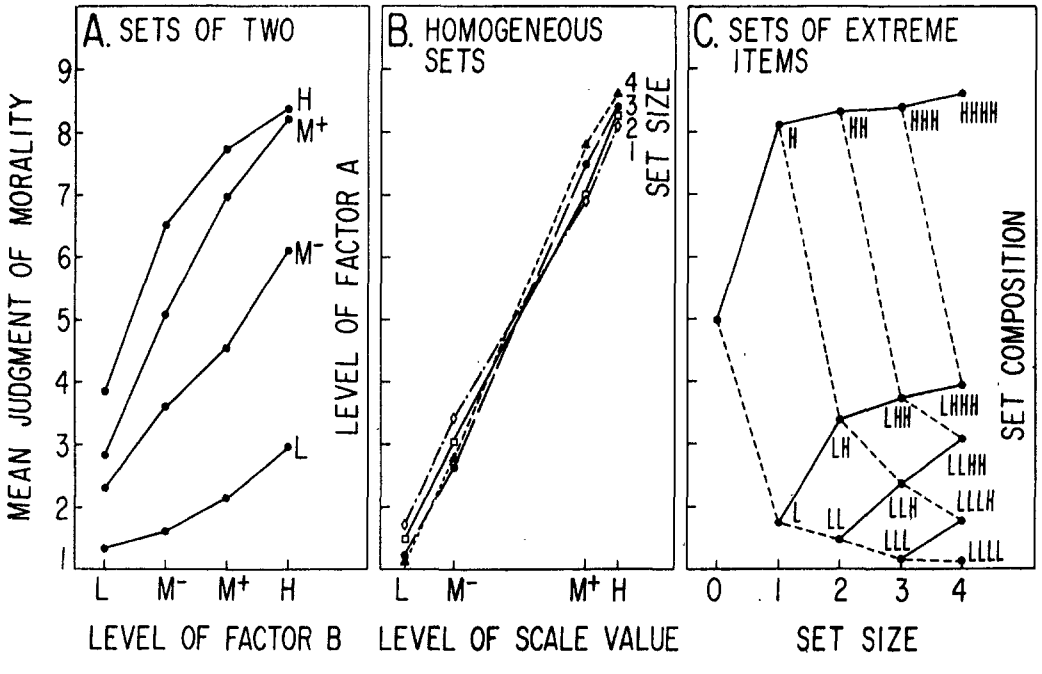

FIG. 1. (A) Mean judgments of morality of pairs of items as a function of the levels of morality (L = low; M^- = moderately negative; M^+ = moderately positive; H = high) of the 2 items. (B) Mean judgments of sets of items of homogeneous scale value as a joint function of scale value and set size. (C) Mean judgments of sets containing extremely moral (H) and extremely immoral (L) items as a function of the number of items in the set and set composition.

Figure 1C plots mean judgments as a function of the number of items and the composition of the set. The upward-sloping solid lines represent the effect of adding a highly moral (H) item; the dashed lines sloping downward represent the effect of adding a highly immoral (L) item. For example, the uppermost solid curve shows the judgments of homogeneous sets as a function of the number of H items; the second solid curve from the top shows the effect of adding additional H items to a set containing a single L item. According to the averaging model, both curves should approach asymptote at the same limit. Instead, the curves suggest that it would take many good acts to undo the effect of just 1 bad one; perhaps no amount of good would make the overall impression favorable. This configural effect is inconsistent with the general averaging model which requires that the addition of H items should raise the rating toward a limit at the value of H.

Discussion: Model Analysis

According to the averaging model with differential weights, the overall impression of morality, Ψ, of k items is given by the equation

$$\Psi = [w_0 s_0 + \sum_{i=1}^{k} w_i s_i]/[w_0 + \sum_{i=1}^{k} w_i], \quad [1]$$

where w_i and s_i are the weight and scale value of the ith level of morality and w_0 and s_0 are the weight and scale value of the initial impression. The value of the initial impression can be estimated from the data and usually corresponds to the neutral point on the response scale. Without loss of generality, the origin of the scale values can be defined by setting $s_0 = 0$, and the unit of the weights can be defined by setting $w_0 = 1$.

The number of items, k, is called the set size. When the scale values of the items in the set are held constant, the set is termed "homogeneous." The effect of set size for homogeneous sets derives from Equation 1:

$$\Psi = s_i[kw_i/(w_0 + kw_i)], \quad [2]$$

where Ψ is the impression of k items, all of scale value s_i, with weight w_i.

Constant-weight averaging and additive models assume that the weights are independent of scale value. These models reduce to multiplicative functions of set size and scale value. For the additive model, $\Psi = skw$; for the constant-weight averaging model with initial impression, $\Psi = s \cdot [kw/(w_0 + kw)]$.

Manis et al. (1966) have suggested a somewhat different model in which set size has an effect proportional to the logarithm of the set size plus 1: $\Psi = s \cdot \log(k + 1)$.

Each of these models is a multiplicative function of set size (k) and scale value (s). Therefore, all of these models predict that the interaction between these 2 factors should be located entirely in the bilinear component (Anderson, 1970). In this respect, their implications differ from those of the averaging model with differential weights (Equation 2), which can account for the residual from the bilinear interaction shown in Figure 1B by assuming greater weight for more extreme high or low stimuli. The greater the weight of a stimulus level, the less the effect of the initial impression on judgments of sets of these stimuli; therefore, the effects of set size are relatively less when the items in the set have greater weight.

Since the range model reduces to a constant-weight averaging model when the items are all of the same value, it cannot account for the residual from bilinearity in Figure 1B. The range model also predicts a steady divergence for Figure 1A. Although the general trend is one of divergence, the top curve in Figure 1A shows a slight reconvergence at the upper end.

It is useful to consider the effects of transforming the data so that the curves in Figure 1B are linear. This mild transformation would have the effect of stretching the scale at the ends and would make the data in Figure 1A in better agreement with the range model. However, such data transformation may not be appropriate. Further constraints would be required to determine whether the slight reconvergence in Figure 1A is "real" or due to a slight nonlinearity in the rating scale (see Birnbaum, 1972b).

The differential-weight averaging model can account for reconvergence in Figure 1A and for residual from bilinearity in Figure 1B when fit to either set of data separately. However, the averaging model with differential weights cannot account for the data of Figure 1C. A least squares solution for the weights and scale values yields predictions that are too steep for the L-LH-LHH-LHHH curve and too flat for the L-LL-LLL-LLLL curve. The averaging model predicts that the L-LH-LHH-

LHHH curve should have the same asymptote as the H-HH-HHH-HHHH curve.

There are several possible interpretations of this configural effect. The change-of-value interpretation assumes that the H item receives a lower value in the context of a single L item. Having committed an L deed, a person's H deeds are perceived as slightly immoral. Thus, the L-LH-LHH-LHHH curve asymptotes at the configural value for H. However, it seems unlikely that component ratings of the individual H deeds would support this change-of-value interpretation. The *S* would probably report that the H deeds are themselves good, but the *person* who has committed an L deed must be bad.

A second interpretation assumes that the L item sets an upper bound on the overall impression. Having committed an L deed, the person cannot be considered moral, no matter how many good deeds he performs. This interpretation predicts that the asymptote depends upon the value of the L item and would be independent of the value of the H item. Thus, a curve consisting of L-LM^{+}-LM^{+}M^{+}- . . . should approach the same limit as the L-LH-LHH-LHHH curve.

A third possibility is that the weight of an item depends in part upon its rank within the set. For example, the overall impression may be a weighted average of the *worst* deed and the average of the others. This configural-weight model is a more general case of the range model that would require specification of the configural weights.

In summary, this research indicates that morality judgment may represent a truly configural process. Overall impressions of morality do not appear to be simple weighted averages of the separate values of a person's deeds. The present data suggest instead that bad deeds have an overriding impact on the overall judgment. A person may be judged mostly by his worst bad deed.

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. Averaging model analysis of set size effect in impression formation. *Journal of Experimental Psychology*, 1967, **75**, 158–165.

ANDERSON, N. H. A simple model for information integration. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally, 1968.

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

ANDERSON, N. H. Looking for configurality in clinical judgment. *Psychological Bulletin*, 1972, **78**, 93–102.

BIRNBAUM, M. H. Morality judgments: Tests of an averaging model. *Journal of Experimental Psychology*, 1972, **93**, 35–42. (a)

BIRNBAUM, M. H. The nonadditivity of impressions. Unpublished doctoral dissertation, University of California, Los Angeles, 1972. (b)

BIRNBAUM, M. H., PARDUCCI, A., & GIFFORD, R. K. Contextual effects in information integration. *Journal of Experimental Psychology*, 1971, **88**, 158–170.

FISHBEIN, M., & HUNTER, R. Summation versus balance in attitude organization and change. *Journal of Abnormal and Social Psychology*, 1964, **69**, 505–510.

MCGARVEY, H. R. Anchoring effects in the absolute judgment of verbal materials. *Archives of Psychology*, 1943, **39**, No. 281.

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.

(Received August 24, 1972)