JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY 19, 157–171 (1983)

Contextual Effects in Social Judgment

BARBARA A. MELLERS

University of California, Berkeley

AND

MICHAEL H. BIRNBAUM

University of Illinois, Champaign

Received July 14, 1981

Judges were asked to evaluate the overall performance of hypothetical students, given their scores on two examinations. The distribution of total scores was manipulated in order to investigate the loci of contextual effects. The interaction between the two exams was reversed by manipulation of the distribution. When the distribution of total scores was positively skewed, judgments showed a convergent interaction as a function of the two exams; when the distribution was negatively skewed, the interaction was divergent. The data were consistent with the hypothesis that the distribution of total scores affects only the transformation from integrated impressions to overt responses. This transformation (judgment function) was well-described by an extension of range-frequency theory. The finding that the interaction can be manipulated by changing the stimulus distribution has methodological implications for the popular interpretation of interactions or lack thereof. A good model may be improperly rejected or a bad one improperly retained through lack of attention to contextual effects.

Many social judgments such as jury decisions, evaluations of applicants, and likableness judgments require the integration of two or more pieces of information. To investigate models of how the subject integrates information, judgments are often plotted as a function of the value of one cue with a separate curve for each level of the other cues. This graph is then interpreted as evidence for a particular model of information integration. If either an additive or parallel-averaging model is appropriate

This research was supported in part by funds from Research Board, University of Illinois. Coauthorship is equal. Requests for reprints may be addressed to either author: Barbara A. Mellers, Department of Psychology, University of California, Berkeley, CA 94720; Michael H. Birnbaum, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820.

0022-1031/83/020157-15$03.00/0

and if the dependent variable is assumed to be a linear function of subjective value, the curves in the graph should be parallel (Anderson, 1979; Birnbaum, 1974a, 1982). Parallel data are usually interpreted as consistent with an additive or averaging model, whereas nonparallel data are taken as evidence against additive or parallel-averaging models. Such interactions are often interpreted as evidence for multiplying, differentially weighted averaging, or configural-weight averaging models.

The present paper warns that the parallelism test cannot be interpreted so easily: interactions should not necessarily be attributed to the combination or comparison rule. In particular, this paper extends developments of Birnbaum, Parducci, and Gifford (1971) and Mellers and Birnbaum (1982) to show that the stimulus distribution affects the transformation of subjective values to overt responses according to range-frequency theory. Interactions can be made to occur, and be reversed, by varying the distribution of the stimuli presented for judgment.

Previous research using judgments of line length found evidence consistent with the hypothesis that the distribution of integrated impressions affects the judgment function (Birnbaum et al., 1971). The present research is designed to investigate contextual effects produced by manipulation of the joint stimulus distribution using a social judgment task and to relate these effects to judgments of single items. The subject's task is to evaluate the overall performance of hypothetical students as a function of the scores on either one or two exams. The joint distribution of the two exam scores is manipulated to vary the distribution of integrated impressions.

Range-Frequency Theory for Single Ratings

Range-frequency theory was developed by Parducci (1963, 1965, 1974, 1982) to account for ratings of single stimuli presented in different distributions. In Birnbaum's (1974b) notation, the theory can be written as

$$G^*_{ik} = J^*_k(s_i) \tag{1}$$

where G^*_{ik} is the category rating of stimulus i in context k, J^*_k is the judgment function in context k, and s_i is the subjective value of stimulus i. Range-frequency theory specifies that the J^*_k function represents a compromise between two tendencies: (a) a tendency to use equal portions of the response range for equal portions of the subjective stimulus range, and (b) a tendency to use equal portions of the response range with equal frequency. When the stimulus range is held constant, the model can be written

$$G^*_{ik} = a^*_k\{\alpha^* F^*_k(s_i) + (1-\alpha^*)s_i\} + b^*_k \tag{2}$$

where $F^*_k(s_i)$ is the frequency component (the cumulative proportion of stimuli having subjective values less than s_i in context k); s_i is the range component (scale value on a zero to one scale); α^* is the weight of the frequency component. The expression inside the brackets varies from 0 to 1; a_k^* and b_k^* are linear constants that convert the predictions to the appropriate response scale.

Theories of Contextual Effects for Combination Ratings

Range-frequency theory has been extended by Birnbaum et al. (1971), Mellers and Birnbaum (1982), and Mellers (1982) to situations involving stimulus comparison and combination. When there are two stimuli to be combined, as in the present study, the context consists of a joint stimulus distribution that has three aspects of theoretical interest: the marginal distributions of the first and second stimuli and the distribution of integrated impressions. In the present study of class performance, there are three distributions of concern: the distribution of scores on the first exam, the distribution of scores on the second exam, and the joint distribution of overall performance. Once the joint distribution is known for context k, the marginals are determined.

Suppose the integrated impression of a student's overall performance is an additive (or parallel-averaging) combination of separate impressions of the student's performances on the two exams

$$\psi_{ijk} = s_{ik} + t_{jk} \tag{3}$$

where ψ_{ijk} is the impression of overall performance of a student with exam performances i and j in context k; s_{ik} and t_{jk} are the separate impressions of exam scores i and j on the first and second exams, respectively, in context k; s_{ik} and t_{jk} may depend on the marginal distributions of the first and second exams, respectively. In the present study, marginal distributions are the same within each context; therefore, it is assumed that $s_{ik} = t_{ik}$.

A general theory of contextual effects in judgments of overall performance can be written

$$G_{ijk} = J_k(s_{ik} + t_{jk}) \tag{4}$$

where G_{ijk} is the rating of a student with exam levels i and j in context k; J_k is the judgment function for context k; s_{ik} and t_{jk} are the separate impressions that depend on the distributions of scores on the two exams.

The judgment function, J_k, is theorized to depend on the distribution of ψ_{ijk} in context k, according to range-frequency theory. That is, J_k is given by Eq. 2, substituting ψ_{ijk} for s_i. The F_k would represent the cumulative density function for ψ_{ij} in context k. Hence, range-frequency theory is being extended from a theory of J_k^* for single stimulus judgments to a theory of J_k for multiple stimulus judgments.

Special cases: two theories. A special case of Eq. 4 assumes that the judge considers each item score in the corresponding distribution of scores for the exam and *then* combines the separate impressions to arrive at a judgment of overall performance. This theory can be written

$$G_{ijk} = J(s_{ik} + t_{jk}) \tag{5}$$

where the judgment function is the same for all contexts, but the separate impressions s_{ik} and t_{jk} depend on context, as in Eq. 2; (i.e., $s_{ik} = G_{ik}$).

A second theory, which is also a special case of Eq. 4, assumes that contextual effects operate only on the distribution of integrated impressions, that is, the sum of the two exams. This theory can be expressed as

$$G_{ijk} = J_k(s_i + t_j) \tag{6}$$

where the J_k function depends on the context, but the scale values, s_i and t_j, do not. This theory assumes that the overall performance judgment depends on the relative position of a student's total exam score in the distribution of total exams (or overall performance). Evidence for this special case (Eq. 6) was found by Birnbaum, Parducci, and Gifford (1971) for psychophysical judgments.

Null hypothesis. If the context has no effect or if the context has only linear effects on the scale values and/or judgment functions, then judgments will be linearly related across contexts. Such linear theories of contextual effects are implied by adaptation-level theory and correlation–regression theory (see Birnbaum, 1974b).

Distinguishing among the theories. The general theory of Eq. 4 implies that in an experiment manipulating the joint distribution of Exam 1 and Exam 2, plots of judgments as a function of levels of Exam 1 with a separate curve for each level of Exam 2 will be systematically nonparallel due to the effect of J_k. The shape of the curves should differ for different distributions of totals because J_k differs (i.e., convergent with a positively skewed distribution and divergent with a negatively skewed distribution). Furthermore, the separate impressions of each exam for the additive model, s_{ik} and t_{jk}, should vary systematically with the marginal stimulus distribution (i.e., the distribution of each single exam) according to range-frequency theory. Because the rank orders of the judgments will change with nonlinear transformations of the scale values, the rank orders will vary for different contexts.

Equation 5 implies that within each contextual condition responses will be an additive function of the separate impressions s_{ik} and the monotonic transformation that renders the curves parallel (J^{-1}) will be the same for all conditions. However, the rank orders of overall performance will differ since they depend on the separate impressions s_{ik} which are context dependent.

Equation 6 implies that the curves will be systematically nonparallel due to the effect of J_k. Furthermore, the nonparallelism will differ for different conditions. This model also implies that rank orders of the overall performance judgments will be the same irrespective of the joint distributions, because the scale values are assumed to be independent of context.

METHOD

The subject's task was to evaluate the overall performance of hypothetical students on the basis of their scores on either one or two exams. Different groups of judges received one of two distributions of exams (representing hypothetical classes) in which the students' total exam scores were either negatively or positively skewed.

Instructions

In the conditions with two exams, judges were told to assume the tests were comparable and equal in length. Judges were asked to rate each hypothetical student by recording a number which represented the student's overall performance. The following response scale was provided for all of the conditions: 9, Very Very Good performance; 8, Very Good performance; 7, Good performance; 6, Slightly Better Than Average performance; 5, Average performance; 4, Slightly Worse Than Average performance; 3, Bad performance; 2, Very Bad performance; 1, Very Very Bad performance.

Design

Figure 1 illustrates the bivariate distribution for the positively skewed condition. Scores on Exam 1 and 2 are plotted on the abscissa and ordinate, respectively. Each square, circle, and triangle represents the performance of one, one, or three hypothetical students, respectively, on the two exams.

Squares in Fig. 1 represent the pairs of test scores common to both the positively and negatively skewed distributions of total exam scores. For example, the square in the lower, left-hand corner represents a hypothetical student who scored 5 on both exams. These 40 common stimuli were constructed from the union of two overlapping factorial designs of Exam 1 by Exam 2 (note the 4 × 7 design and the 7 × 4 design in the squares shown in Fig. 1). In addition, each distribution had 120 context stimuli, shown as triangles and circles for the positively skewed condition in Fig. 1.

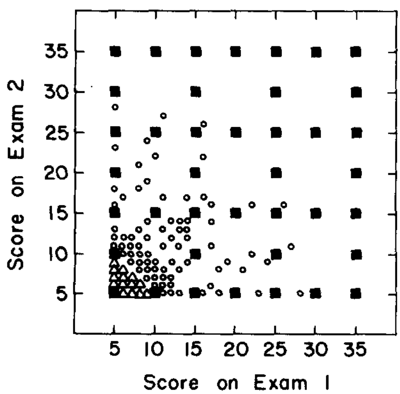

FIG. 1. Joint distribution of exam scores for positively skewed condition. Each circle or square represents the performance of one hypothetical student. Each triangle represents the performance of three students with the same scores. Squares are the stimulus trials common to both contexts. Contextual trials for negatively skewed distribution were the mirror image of those in this figure, reflected about the axis, Exam 1 + Exam 2 = 40.

The distribution for the negatively skewed distribution of total exams can be obtained from Fig. 1 by reflecting the open symbols about the diagonal axis defined by the line, Exam 1 + Exam 2 = 40. In the positively skewed distribution of total scores, the marginal exam distributions of Test 1 and Test 2 were both positively skewed; in the negatively skewed distribution of totals, the marginal distributions were both negatively skewed.

Conditions

There were six experimental conditions, and each judge served in one of them. In two conditions, subjects were given either the positively skewed trials or negatively skewed trials and were asked to judge performance on the basis of two exam scores. In two other conditions, judges were given the same information plus a histogram showing the distribution of scores for Exam 1. They were told that the distribution of Exam 2 was exactly the same, although a given student may not have obtained the same score on both exams. It was thought that the histogram might enhance a possible tendency to compare each score with its distribution and then combine two judgments of performance (as in Eqs. 4 and 5). In two more conditions without histograms, subjects were asked to evaluate students in either the negatively or positively skewed distributions on the basis of only one exam (Exam 1). The distribution of test scores on Exam 1, for the positively skewed condition for example, can be determined by projecting the points in Fig. 1 on to the abscissa.

Procedure

Subjects were given 2 pages of instructions, 1 page with 15 representative warm-up trials, and 3 pages of experimental trials. The warmups (of either single exam scores or pairs of exams depending on the condition) were selected to familiarize them with the range and relative frequency of the test scores. Each trial consisted of one or two numerical values. Trials were presented in a random order, and page order was varied across subjects. Subjects were permitted to work at their own paces; most of them completed the task in about 45 min.

Subjects

Subjects were 160 undergraduates at the University of Illinois, Urbana-Champaign, who received extra credit in lower division psychology classes for participating. There were between 23 and 30 different judges in each of the six conditions. Judges were assigned to conditions by a computer program that scheduled sessions. In each session, 5 to 15 judges were each given booklets for the same condition, but they worked independently. An additional 3 people were tested who failed to follow instructions and whose data were excluded.

RESULTS

The mean judgments for the common trials are plotted in Fig. 2 as a function of the score on Exam 1 with a separate curve (and symbol) for each level of Exam 2. Judgments of the common stimuli are higher for the positively skewed context, where most of the students had low totals, than for the negatively skewed context, where most of the students had higher totals.

Parallelism would be consistent with an additive (or parallel-averging) model, assuming a linear J_k function. Instead of being parallel, however, the curves for the positively skewed context show systematic convergence to the right (the vertical separations between the curves decrease as the

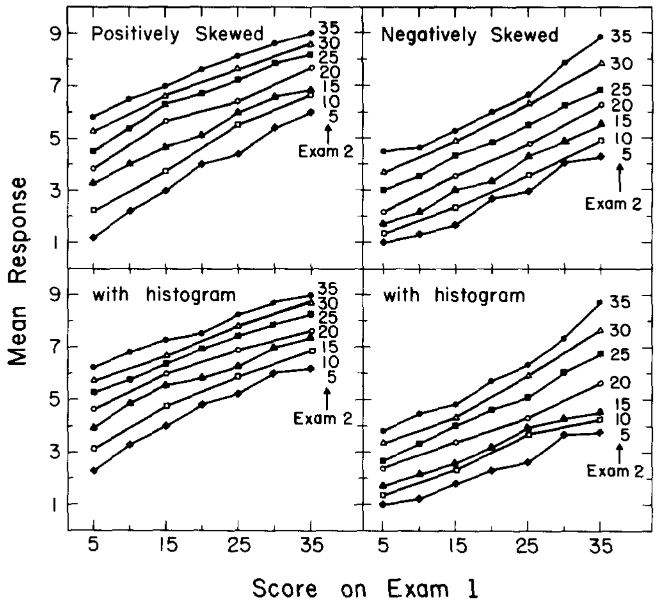

FIG. 2. Mean evaluations of the performance of students as a function of score on Exam 1 with a separate curve and separate symbol for each level of score on Exam 2. Note that the curves converge when the exam totals are positively skewed, and they diverge when the totals are negatively skewed.

score on Exam 1 increases). On the other hand, curves for the negatively skewed context show divergence to the right. The interaction between exam scores was statistically significant in all four panels of Fig. 2.[1]

The shape of interaction changes from convergent to divergent for positively skewed versus negatively skewed contexts. The three-way interaction of Context by Exam 1 by Exam 2 was statistically significant. However, this interaction did not appear to depend on the histogram versus no histogram manipulation.[2] Because the effects of the histogram versus no histogram manipulation were minimal, this factor is disregarded in subsequent analyses.

Individual subject analyses revealed that the means in Fig. 2 were

[1] The first 23 subjects were used from each of the 4 conditions for analyses of variance. Eight separate ANOVAs yielded the following F's for the Exam 1 by Exam 2 interactions for the 7×4 and the 4×7 subdesigns, respectively: Positively skewed with histogram, $F(18, 396) = 3.62$ and 3.39; positively skewed without histogram, $F = 2.95$ and 3.76; negatively skewed with histogram, $F = 5.53$ and 7.93; negatively skewed without histogram, $F = 2.23$ and 2.62. The critical value of $F(18, 396)$ at the .01 level is about 2.0.

[2] The three-way interaction of Context (Positive vs Negative) by Exam 1 by Exam 2 yielded $F(18, 1584) = 8.96$ and 11.64 for the 7×4 and the 4×7 subdesigns, respectively, averaged over the histogram manipulation. The four-way interaction of Histogram by Context by Exam 1 by Exam 2 yielded $F(18, 1584) = 1.15$ and 1.42 for the 7×4 and 4×7 subdesigns, respectively.

representative of the majority of individual subjects. Differences between the highest and lowest curves were computed for two values—5 and 35 on the abscissa (Exam 1). For 39 out of 52 subjects in the two positively skewed contexts (histogram and no histogram), the difference between the curves is greater at the value of 5 than 35 (i.e., convergence). The curves for the two negatively skewed contexts show the opposite: divergence to the right. Thirty-seven out of fifty-four subjects in these contexts showed this divergent interaction in their judgments.

Locus of Contextual Effects

Because the interaction changes from convergent to divergent, the data are inconsistent with Eq. 5, which assumes that J functions are independent of context and therefore requires the same interaction in all panels of Fig. 2. This result also rules out the null hypothesis that the contextual effects are linear.

Further evidence that the contextual effects are nonlinear is shown in Fig. 3A. Marginal means for the positively skewed (two-exam) context are a negatively accelerated function of marginal means for the negatively skewed (two-exam) context. However, marginal means are not scale values, and this nonlinear relationship does not necessarily imply that scale values changed as a function of context. This nonlinear relationship is consistent with both special cases of Eq. 4.

Equation 4 was fit to the data in each panel of Fig. 2 separately by means of MONANOVA (Kruskal & Carmone, 1969). A different J_k function was estimated for each of the four conditions, and different s_{ik} and t_{jk} values were estimated for Exam 1 and Exam 2 in each condition. It was

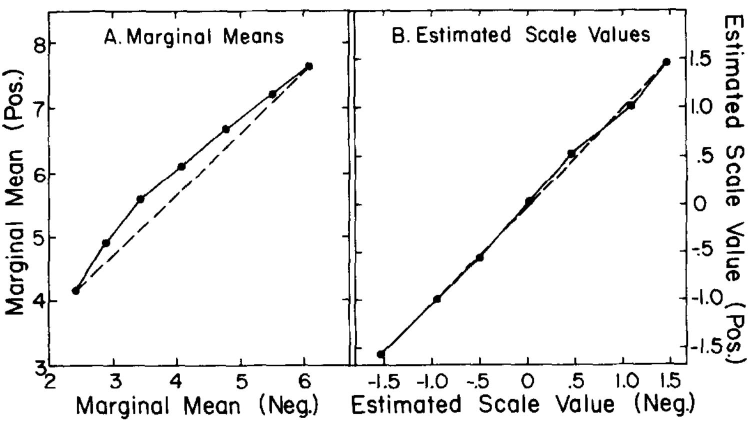

FIG. 3. A comparison of the marginal means and the estimated scale values (marginal means after rescaling to additivity) for the common stimuli in the two contexts based on the two-exam judgments. Note that the marginal means in the two conditions are nonlinearly related, but that estimated scale values in the two conditions are nearly identical.

found that estimated scale values were virtually identical for first and second exam scores ($s_{ik} = t_{ik}$), which was expected because the two exams had the same marginal distributions within each condition. Furthermore, it was found that estimated scale values were independent of positive skew versus negative skew context and histogram versus no histogram. Estimated scale values for the positively skewed condition (averaged over first vs second exam and histogram vs no histogram) are shown as a function of estimated scale values for the negatively skewed context in Fig. 3B. The dashed line connects the end points, and it is clear that the data fall very close to the line.

The finding that scale values estimated from Eq. 4 are virtually identical for the positively and negatively skewed contexts (Fig. 3B) suggests that Eq. 6 can be used to account for the data. Thus, although marginal means are nonlinearly related across contexts, (Fig. 3A); the data are consistent with the hypothesis that the model and scale values are the same across contexts and only the judgment function changes.

Fit of Range-Frequency Theory to the Judgment Function

Two-exam judgments. Manipulation of the joint stimulus distribution appears to influence the judgment function relating integrated impressions ψ to overt responses G. Range-frequency theory was fit to the judgment function in two ways. The first analysis assumed the additive model and used the average scale value estimates of Eq. 4 to compute ψ_{ijk}. The second method estimated the values of ψ from the data using range-frequency theory in order to provide a check on the assumption of additivity.

The eight estimates of each scale value from Eq. 4 were averaged and used to compute ψ_{ijk} values for the 40 common trials (squares in Fig. 1). It was assumed that $\psi_{ijk} = s_i + s_j = \psi_{ij}$ for all k, according to Eq. 6. Mean judgments for the common stimuli from Fig. 2 (averaged over histogram and no histogram) were linearly recalibrated to make the extreme judgments 0 and 1. These judgments are plotted as a function of the ψ_{ij} values in Fig. 4, using the same symbols as in Fig. 2. A straight line has been drawn through the extreme judgments, leaving 38 points from each context free to vary.

All of the 38 symbols above the straight line in Fig. 4 are from the positively skewed context and all of the 38 symbols below the line are from the negatively skewed context. If there had been no contextual effects or if all contextual effects had been linear, then all of the symbols would have fallen on the same function in Fig. 4. Equation 6 implies that all of the points within each context fall on a monotonic function, with a different monotonic function for each context. Furthermore, if range-frequency theory describes the J_k functions of Eq. 6, judgments should be the average of the solid curve for that context (frequency

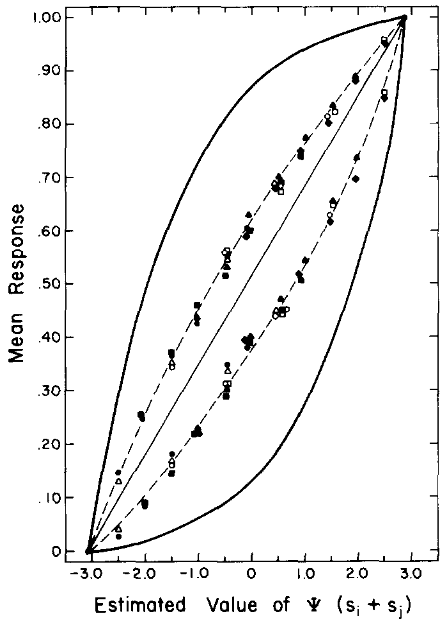

FIG. 4. Mean judgments of student performance based on two exams (recalibrated to a 0–1 scale) as a function of estimated values of $\psi = s_i + s_j$. Symbols correspond to those used in Fig. 2. Solid curves are cumulative density functions for the two contexts; dashed curves are predictions of range-frequency theory.

component) and the straight line through the endpoints (range component). The solid curves show the two cumulative density functions for the distributions of total exam scores. The dashed lines are the least-squares estimates of the fit of range-frequency theory to the data in Fig. 4. There appears to be no pattern to the small deviations in Fig. 4.

The second analysis assumes range-frequency theory as a theory of the J_k function and uses it to solve for values of ψ_{ij} to test the additive model. The following model was fit to the data:

$$G_{ijk} = \alpha F_k(\psi_{ij}) + \psi_{ij} \tag{7}$$

where G_{ijk} is the overall performance rating of a student with exam scores i and j in context k; $F_k(\psi_{ij})$ is the cumulative proportion of students in context k receiving lower total exam scores (solid curves in Fig. 4); ψ_{ij} are parameters estimated from the data (which are not required to be additive); α is a fitted constant. This analysis yielded a multiple correlation of .999, compared with .998 for the analysis that assumed additivity. The estimated ψ_{ij} values were very nearly parallel (with a very small divergence). The additive or parallel-averaging model appears to provide a satisfactory approximation in this case.

One-exam judgments. Figure 5, plotted as in Fig. 4, shows data (points)

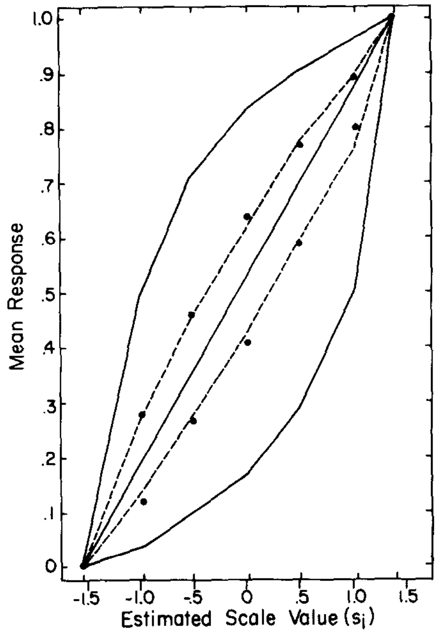

FIG. 5. Mean judgments of student performance based on one exam (recalibrated to a 0–1 scale) as a function of estimated value of s_i. Solid curves are cumulative density functions for the two contexts; dashed curves are the predictions of range-frequency theory.

and predictions (dashed lines) of range-frequency theory for the two conditions with judgments based on only one exam. Mean judgments were linearly recalibrated to a 0 to 1 scale for each context. The solid curves show the cumulative density functions for Exam 1 in the positively skewed context (upper curve) and negatively skewed context (lower curve). Averaged scale values estimated from Eq. 4 using the two-exam judgments are plotted on the abscissa. The solid lines connect the endpoints. The dashed lines give a good approximation of the single-exam judgments.

DISCUSSION

Theoretical Implications

The data appear consistent with the following premises.

(1) The effect of variation of the stimulus distribution could be attributed to changes in the judgment transformation from subjective impressions to overt responses, that is, the J_k function

$$G_{ijk} = J_k(\psi_{ij}) \tag{8}$$

$$G^*_{ik} = J^*_k(s_i) \tag{9}$$

where G and G^* are two-exam and one-exam judgments, respectively, and J_k and J^*_k are the respective judgment functions.

(2) Range-frequency theory could describe the judgment functions for both the two-exam and one-exam conditions. For two-exam conditions, the distribution of totals determines F_k; for one-exam conditions, the marginal distribution of each exam determines F_k^*

$$J_k(\psi_{ij}) = a_k\{\alpha F_k(\psi_{ij}) + (1-\alpha)\psi_{ij}\} + b_k \quad (10)$$
$$J_k^*(s_i) = a_k^*\{\alpha^* F_k^*(s_i) + (1-\alpha^*)s_i\} + b_k^* \quad (11)$$

where ψ and s have been scaled from 0 to 1.

(3) The additive (or parallel-averaging) model gave a good approximation to the combination of separate impressions to form an overall evaluation

$$\psi_{ij} = s_i + s_j. \quad (12)$$

(4) The scale values for levels of exam performance were independent of the following variables: first or second exam, one-exam or two-exam tasks, histogram or no histogram, and positively or negatively skewed contexts. (Note that in the above premises, there are no subscripts for context (k) on s or ψ, and no other subscripts representing other variables. Thus, the same scale values were used for one-exam and two-exam tasks.)

These premises can explain the findings that the interaction between the two exams can be manipulated (Fig. 2), that marginal means are nonlinearly related across contexts (Fig. 3A), but scale values are independent of context (Fig. 3B), and that both two-exam and one exam judgments are well-fit by range-frequency theory (Fig. 4, 5).

Related Research on Contextual Effects

The present results are compatible with the results of Birnbaum et al. (1971) and Mellers and Birnbaum (1982, Experiment 2) who concluded that when stimuli from the same modality are compared, the scale values are independent of the stimulus distribution and the procedure for responding. However, related research (Mellers & Birnbaum, 1982, Experiment 3; Mellers, 1982, Experiment 4) indicates that when stimuli are compared or combined across *different* modalities, the scale values depend on the marginal distribution according to range-frequency theory in each modality. For example, Mellers (1982) obtained inequity judgments of faculty members as a function of their salaries and merits. Both the joint distribution of salaries and merits and the marginals were varied. Mellers concluded that inequity judgments are a type of cross-modality comparison in which the scale value of salary depends on the distribution of salaries, the scale value of merit depends on the distribution of merits, and the judgment of inequity depends on the difference between the salary and merit scale values and the distribution of differences.[3]

[3] In the present study, the marginal distributions of Exam 1 and Exam 2 were identical, and no effect of context was found on scale values. These results seem analogous to the

Manipulating the Interaction

The present study demonstrates that the interaction can be manipulated by varying the stimulus distribution. According to Birnbaum's (1974b, Eq. 10) range-frequency analysis, it should also be possible to manipulate the J_k function by varying the response distribution.

Surber (1981) asked judges to predict performance on a hypothetical exam as a function of the student's IQ and study time. It was found that interaction between IQ and study time could be reversed by varying the difficulty of the exam. When the exam was described as "easy" (producing a negatively skewed distribution of performance scores) performance judgments showed a convergent interaction: exam scores were judged to be high when either IQ or study time was high. When the exam was said to be "difficult" (with a positively skewed distribution of performance scores), the interaction between IQ and study time was divergent: performance was high only when both variables were high. Surber (1981) discussed several interpretations of her data including the possibility that exam difficulty affects the J function according to range-frequency theory.

It is also possible to manipulate the interaction in ways that cannot be explained by the judgment function. Birnbaum and Stegner (1979) obtained judgments of the value of used cars as a function of estimates provided by sources who examined the cars. They found that the interaction between two estimates of the value of a used car could be manipulated by changing the subject's point of view from that of a buyer to that of a seller. Since the data show consistent changes in rank order (see Birnbaum, 1982), their results cannot be explained by the theory that point of view affects only the J function.

Methodological Implications

Parallelism test. These results show that contextual effects could alter the conclusions of studies that use the parallelism test to evaluate theories. Although the curves in Fig. 2 were not parallel, an additive (or parallel-averaging) model was retained for the data.

It should also be possible to select stimuli to produce parallelism, even though the combination rule is *not* an additive one. Some authors have advocated the use of extreme end anchors, filler stimuli, graphic rating scales, and other procedures based on the contention that these methods ensure a linear J function. However, from the present viewpoint, extreme stimuli, fillers, etc. can affect the J function but do not necessarily make it linear. For example, to predict the effect of end anchors on the J

---

within-mode research of Mellers and Birnbaum (1982, Exp. 2). However, if the two exams had very different distributions, the task might be cross modal, and the scale values might depend on the context.

function requires prior knowledge (or experimental manipulation) of the subjective values of the end stimuli and of the distribution of subjective values.

When the scale values and model are unknown, the distribution of ψ is unknown. If the true model produces a divergent interaction, then the distribution of ψ could be positively skewed. Range-frequency theory implies that a positively skewed distribution will induce a negatively accelerated J function, which tends to produce *convergence.* Depending on the scale values, the value of α, etc., it is possible that the data will appear parallel. Thus, the test of parallelism in scale-dependent research may be systematically biased (See Birnbaum (1982) for further discussion).

Marginal means and single ratings. A further methodological implication of the present data is that neither single ratings nor marginal means should necessarily be regarded as scale values. Because the marginal means and the single ratings are both nonlinearly related across contexts, they would predict *two* different rank orders in the two-exam conditions, neither of which is consistent with the rank order of the judgments.

A study by Brehmer and Slovic (1980) illustrates the traditional interpretation of single ratings and marginal means of combinations. They found approximate linearity between marginal means and single ratings and concluded that (1) scale values did not change in the integration process, and (2) single ratings were good estimates of scale values. However, the present results show that the relationship between marginal means, scale values, and single ratings is not linear in general and depends on the stimulus distribution. Agreement between marginal means and single ratings does not constitute a diagnostic test of linearity and additivity; nor would nonlinearity imply that the scale values had changed.

REFERENCES

Anderson, N. H. Algebraic rules in psychological measurement. *American Scientist*, 1979, **67**, 555–563.

Birnbaum, M. H. The nonadditivity of personality impressions. *Journal of Experimental Psychology*, 1974, **102**, 543–561. (a)

Birnbaum, M. H. Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, 1974, **15**, 89–96. (b)

Birnbaum, M. H. Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J.: Erlbaum, 1982.

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

Birnbaum, M. H., & Stegner, S. E. Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, 1979, **37**, 48–74.

Birnbaum, M. H., & Veit, C. T. Scale convergence as a criterion for rescaling: Information integration for difference, ratio, and averaging tasks. *Perception & Psychophysics*, 1974, **15**, 7–15.

Brehmer, B., & Slovic, P. Information integration in multiple cue judgments. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, **6**, 302–308.

Kruskal, J. B., & Carmone, F. J. MONANOVA—A FORTRAN IV program for monotone analysis of variance. *Behavioral Science*, 1969, **14**, 165–166.

Mellers, B. A. Equity Judgment: A revision of Aristotelian views. *Journal of Experimental Psychology: General*, 1982, **111**, 242–270.

Mellers, B. A., & Birnbaum, M. H. Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance*, 1982, **8**, 582–601.

Parducci, A. Range-frequency compromise in judgment. *Psychological Monographs*, 1963, **77**, (2, No. 565).

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, **72**, 407–418.

Parducci, A. Contextual effects: A range-frequency analysis. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press, 1974.

Parducci, A. Category ratings: Still more contextual effects. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J.: Erlbaum, 1982.

Rose, B. J., & Birnbaum, M. H. Judgments of differences and ratios of numerals. *Perception & Psychophysics*, 1975, **18**, 194–200.

Surber, C. F. Necessary versus sufficient causal schemata: Attributions for achievement in difficult and easy tasks. *Journal of Experimental Social Psychology*, 1981, **17**, 569–586.