Limitations of Additive Conjoint Scaling Procedures: Detecting Nonadditivity When Additivity Is Known to Be Violated

Thomas E. Nygren
Ohio State University

Two sets of three-outcome gambles were constructed to vary factorially along the factors Amount to Lose, Amount to Win, Probability of Losing, and Probability of Winning. Single stimulus ratings of attractiveness and risk were obtained for each of the constructed gambles from 19 subjects. In addition, paired comparison strength of preference and difference in risk judgments were obtained for a subset of these gambles. Two additive conjoint scaling procedures, Carroll's (1972) MDPREF and Johnson's (1975) NMRG, were used to generate predicted paired comparison preference and risk judgments from the single stimulus ratings for each subject. These predictions were then compared with the observed paired comparison judgments. Results indicated that although the goodness-of-fit measures associated with each of the scaling models indicated that the subject's data were being fit very well by the additive models, additivity among the payoff and probability factors was clearly violated. A procedure for detecting nonadditivity is outlined and illustrated with the data. The limitations of using these additive conjoint scaling procedures as predictive techniques when additivity is violated are shown and their implications are discussed.

In recent years a number of computer-based additive conjoint scaling procedures have been developed for use in applied psychological, marketing, and consumer research (Carroll, 1969, 1972; Johnson, 1975; Kruskal, 1965; Srinivasan & Shocker, 1973a, 1973b). Application of these conjoint scaling methods have generally been restricted to prediction of utilities for multi-factor choice alternatives via the decomposition and scaling of "part-worths," or utilities, of the attribute levels of the factors making up the alternatives (cf. Green & Rao, 1971; Green & Srinivasan, 1978). Although an additive representation has the distinct advantage of being conceptually and computationally simple, its validity in many situations seems questionable. Unfortunately, as Green and Srinivasan (1978) have pointed out, "users of conjoint analysis have generally emphasized predictive validity and regarded explanation as a desirable (but secondary) objective" (p. 104). Of particular concern in the present study is the ability (or lack of ability) of these conjoint analysis procedures for detecting small but consistent deviations for additivity.

The purpose of this paper is twofold: (1) to examine empirically the predictive ability of two additive conjoint scaling procedures for judgments of preference and risk when additivity is known to be violated and (2) to illustrate a procedure for detecting nonadditivity. The conjoint analysis procedures

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 367–383

used here illustrate two general approaches for estimating part-worths for factor attribute levels. These methods are the vector model that underlies Carroll's (1972) MDPREF algorithm and monotone regression, as used in Johnson's (1975) NMRG algorithm.

MDPREF and NMRG

In the MDPREF vector model, the rating of a particular alternative j on a monotonic construct (not necessarily preference) for individual i, s_{ij}, is assumed to be a weighted additive function of the perceived attribute levels of alternative j on some set of relevant dimensions. Formally, the model is

$$\hat{s}_{ij} = \sum_{t=1}^{r} y_{it} \cdot x_{jt} \qquad [1]$$

where

$\hat{s}_{ij}$ is the predicted scale value of stimulus j for individual i,
y_{it} is the importance of dimension t for individual i,
and x_{jt} is the projection of stimulus j on dimension t.

Given the appropriateness of Equation 1, paired comparison difference judgments can then be predicted from single stimulus ratings by

$$\hat{d}_{ijk} = \sum_{t=1}^{r} y_{it} \cdot (x_{jt} - x_{kt}) \qquad [2]$$

Hence, if the $\hat{s}_{ij}$'s represent predicted attractiveness or riskiness judgments, then the $\hat{d}_{ijk}$'s would represent predicted strength of preference or difference in risk judgments between stimuli j and k.

The monotone regression procedure used in NMRG employs a simple but related part-worth function model that can be written as

$$\hat{s}_{ij} = \sum_{t=1}^{r} f_{it}(x_{jt}) \qquad [3]$$

where $f_{it}(x_{jt})$ is individual i's part-worth or psychological scale value for stimulus j on dimension t. The perceived difference between stimuli j and k is then

$$\hat{d}_{ijk} = \sum_{t=1}^{r} [f_{it}(x_{jt}) - f_{it}(x_{kt})] \qquad [4]$$

Although the two models are effectively conceptually the same, the computational algorithms employed by MDPREF and NMRG are quite different. MDPREF uses all subjects' data to estimate a common spatial representation for the stimulus points and fits each subject by a vector in this space, whereas NMRG derives estimates of the part-worths directly for one subject at a time. Given an $N \times n$ matrix $\mathbf{S}$ of judgments of n stimuli for each of N individuals, MDPREF finds an eigenvalue-eigenvector solution based on Eckart-Young (1936) decomposition to obtain an $N \times r$ matrix $\mathbf{Y}$ of coordinates of unit-length subject vectors and an $n \times r$ matrix $\mathbf{X}$ of projections of the n stimuli on r dimensions ($r < n,N$), such that $\hat{\mathbf{S}} = \mathbf{Y} \cdot \mathbf{X}'$ is the best least squares approximation to $\mathbf{S}$.

It is important to note that the derived r stimulus dimensions from MDPREF, even with a suitable rotation, may not necessarily represent the underlying perceptual dimensions used by the subjects. The MDPREF algorithm is an internal scaling procedure that can only potentially reveal these dimensions if they contribute differently to subjects' response judgments. In the limiting case, if the subjects' judgments reflect no individual differences, then a one-dimensional solution would be obtained regardless of the number of perceptual dimensions actually used by the subjects in forming their judgments. In practice, however, this limitation of MDPREF can be minimized because (1) in many empirical situations large individual differences are known to exist and (2) appropriate sampling procedures can enhance the extent of individual differences.

Unlike MDPREF, the monotone regression procedure used in Johnson's NMRG procedure requires the construction of the stimulus set in such a way that the levels of the factors are specified a priori in a factorial design. One begins with a matrix $\mathbf{X}$ of order $n \times p$ in which each row consists of dummy variables coded as "1's" and "0's" indicating the presence or absence of each level of the factors for a given stimulus. For 48 stimuli constructed from a three-factor $4 \times 3 \times 4$ design, each row of $\mathbf{X}$ would consist of $p=11$ dummy variables, three 1's and eight 0's, which are used to specify the stimulus completely. In matrix form the model is then $\hat{\mathbf{S}} = \mathbf{X} \cdot \mathbf{B}$, where vector $\hat{\mathbf{S}}$ contains an individual's predicted scale values for the n stimuli and $\mathbf{B}$ is a vector of length p that contains the part-worth estimates. NMRG begins with arbitrary values in $\mathbf{B}$ and attempts, through an iterative procedure, to find a best set of new values for $\mathbf{B}$ such that the predicted ordering for $\hat{\mathbf{S}} = \mathbf{X} \cdot \mathbf{B}$ is as monotonic as possible with the observed ordering in $\mathbf{S}$.

Although both MDPREF and NMRG assume that the composition rule for the factors is additive, neither can test this assumption except in an indirect predictive or goodness-of-fit sense. The work of Shanteau (1975) and Dawes and Corrigan (1974) with linear models suggests, however, that good predictive ability is not a sufficient condition for assessing the validity of the additive composition rule. Krantz and Tversky (1971) discussed three necessary conditions for additivity. These are simple independence among the factors, double or Luce-Tukey cancellation, and joint independence for all pairs of factors. Holt and Wallsten (1974) later developed a computer program CONJOINT that can test for violations of these conditions in a given set of data. In practice, however, this program is somewhat limited in that (1) a complete (or nearly so) factorial design is necessary for adequate tests; (2) without an explicit error theory it is difficult to assess the significance of the violations of the axioms; and (3) for a given individual, the effect of differential weighting of the factors on violations of the above conditions is not at all clear—particularly when a factor has zero or near zero importance.

A Test of Additivity for Judgments of Risk and Preference

Assume that a set of stimuli are constructed to vary along r factors with $p_1, p_2, \cdots, p_r$ respective levels. Further, assume that the p_j levels of any factor j can be meaningfully ordered by an individual in such a way that the observed rating judgments (e.g., preference or risk) are monotonic over increasing attribute levels of a factor when all other factors are held constant. Two sets of data are collected from each individual: (1) single stimulus ratings or rankings of the $p_1 \times p_2 \times \cdots \times p_r$ stimuli or a sufficient subset to provide an adequate additive scaling solution and (2) paired comparison difference judgments for all or a subset of the possible pairs of stimuli. By using Equations 2 and 4, predicted paired comparison differences can be derived from the scaling solutions and compared with the independently obtained observed paired comparison judgments at the individual subject level.

If the constructed factors are, in fact, those used by the individuals and the levels of these factors are monotonic with the subjects' ratings, then the work of Dawes and Corrigan (1974) would suggest that the fit of the single stimulus data to conjoint scaling models like MDPREF or NMRG would be very good, *even if* the underlying additive model is invalid. Moreover, high overall predictive success for the paired comparison validation procedure would still be expected.

It is argued here that this anomaly can be explained in part by a separation of the paired comparison judgments into two classes—"dominance" and "tradeoff" paired comparisons. The set of dominance paired comparisons consists of the pairs for which each level of the r factors for stimulus a is at least as great as the corresponding level for stimulus b. If simple independence holds among the r factors, then there should be few, if any, errors in prediction from the scaling solutions for the dominance comparisons, regardless of whether additivity holds. Tradeoff comparisons, on the other hand, are the remaining pairs for which one stimulus does not dominate the other across all factors. These are, of course, the very comparisons of relevance to the applied researcher in a predictive sense. Errors in prediction for the tradeoff paired comparisons would negate additivity, since they would indicate violations of double cancellation and/or joint independence for pairs of factors.

For a completely crossed factorial design it is easy to compute and to list the number of comparisons resulting in dominance or tradeoff tests. The number of dominance tests N_D is

$$\sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\cdots\sum_{k=1}^{p_r}\left[(p_1+1-i)\cdot(p_2+1-j)\cdots(p_r+1-k)-1\right] \qquad [5]$$

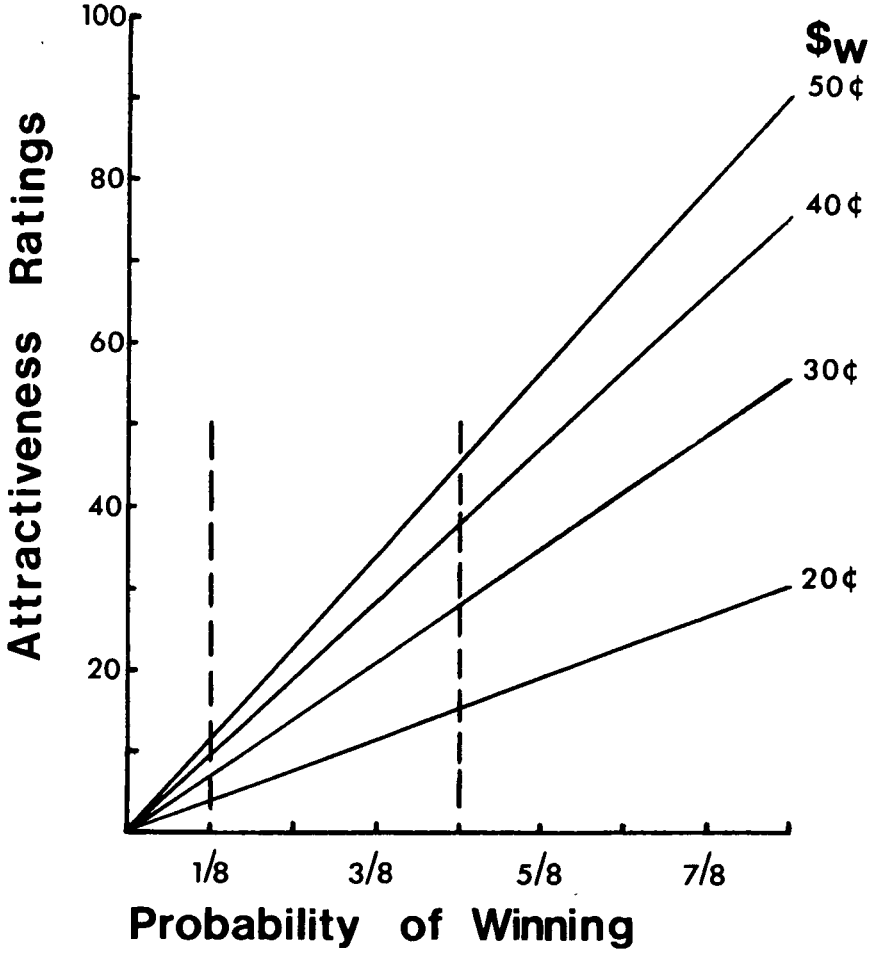
and the number of tradeoff tests is $N_T = N - N_D$, where N is the total number of comparisons. If additivity holds in the data, then under perfect disciminability it would be expected that the proportion of dominance errors, P_D, and tradeoff errors, P_T, would both equal zero. If additivity is violated and independence alone is satisfied, then P_T should be significantly larger than P_D.

The Study

Three-outcome gambles were constructed, each of the form ($-a$¢, 0¢, b¢), where one loses a¢ with probability p_L, wins b¢ with probability p_w, and neither wins nor loses with probability p_0. The gambles were constructed to vary along four factors—Amount to Win ($\$_w$), Amount to Lose ($\$_L$), Probability of Winning (p_w), and Probability of Losing (p_L). These gambles are particularly well-suited to a study of additivity for several reasons. First, each of the four factors above is monotonic with judgments of perceived risk and attractiveness and can be easily varied in a factorial design. Second, there is evidence in the gambling literature that suggests that individual differences are extensive and that each of the varied factors can independently contribute to judgments of perceived risk and attractiveness of gambles (Nygren, 1977; Payne, 1973, 1975; Shanteau, 1974, 1975; Slovic & Lichtenstein, 1968). Finally, the composition rule for risk or attractiveness judgments should *not* be additive (Coombs & Bowen, 1970; Shanteau, 1975).

Figure 1 illustrates a hypothetical example of the interaction effects of the two factors $\$_w$ and p_w, as typically found by Shanteau (1975). On both intuitive and theoretical grounds this interaction between the two factors seems plausible. As either the $\$_w$ or the p_w approach zero, the level of the other factor becomes less relevant. The same is true for $\$_L$ and p_L. Note, however, that if the range of the p_L (or p_w) is constrained, as indicated by the dashed lines in Figure 1, the interaction effect is significantly reduced. Hence, although an interaction model might actually reflect the composition rule employed by an individual in forming judgments of perceived risk or attractiveness, an additive scaling

Figure 1
Hypothetical Illustration of the Interaction Effect between Probability of Winning and Amount to Win for Ratings of the Attractiveness of Gambles on a 100-Point Rating Scale



model would be expected to yield misleadingly excellent results in a goodness-of-fit sense if the ranges of p_L and p_w are appropriately restricted. This implication was tested in the present study for the MDPREF and NMRG models.

In order to reduce the number of judgments made by each subject, two sets of 48 gambles were constructed to vary along three of the factors, $\$_w$, $\$_L$, p_w, and p_L. The designs were $1 \times 4 \times 3 \times 4$ in which $\$_w$ was held constant and $4 \times 1 \times 4 \times 3$ in which $\$_L$ was held constant. The levels of p_w and p_L were restricted to values of 1/8, 2/8, 3/8, and 4/8 in order to reduce possible interaction effects to the small but noticeable levels illustrated in Figure 1. It was hypothesized that this restriction on the probability values would yield MDPREF and NMRG additive scaling solutions with very high fits to the subjects' single stimulus ratings of attractiveness and risk.

Paired comparison judgments of strength of preference and difference in risk were obtained for all pairs of a subset of 16 gambles from each of the larger sets of 48, in a $1 \times 4 \times 2 \times 2$ design and a $4 \times 1 \times 2 \times 2$ design, respectively. The 120 preference strength and risk difference judgments for each set were compared with the predicted orderings generated from Equations 2 and 4 from the scalings of the single stimulus ratings. It was hypothesized that the restrictions on the p_L and p_W values would result in both the MDPREF and NMRG models predicting the observed paired comparison orders significantly better than would be expected by chance. However, when the proportions of dominance (P_D) and tradeoff (P_T) tests were examined, it was hypothesized that $P_D = P_T$ would be consistently violated. It was expected that independence would be substantiated in the data, yielding P_D values close to zero across subjects. However, joint independence should be violated, although moderately because of the restricted ranges of p_L and p_W. These violations should be reflected in consistently larger values of P_T across subjects.

Method

Subjects

The subjects were 20 undergraduate students who were paid at a rate of $2.00 per hour for participating in the experiment. Data from one student was discarded because she failed to follow instructions.

Stimuli

The two sets of 48 gambles constructed from the $1 \times 4 \times 3 \times 4$ (Set A) and $4 \times 1 \times 4 \times 3$ (Set B) designs ($\$_W \times \$_L \times p_W \times p_L$) each used the same probability levels. When three levels of p_W or p_L were used, the values were 2/8, 3/8, and 4/8. When four levels were used, a probability value of 1/8 was added. In the first set of gambles $\$_L$ was either −10¢, −20¢, −30¢, or −40¢ with $\$_W$ constant at 30¢. In the second set $\$_W$ was either 20¢, 30¢, 40¢, or 50¢ with $\$_L$ constant at −30¢. In addition, a subset of 16 gambles were chosen from each of the above sets for use in the paired comparison tasks. The subsets were from a $1 \times 4 \times 2 \times 2$ (Set A′) and a $4 \times 1 \times 2 \times 2$ (Set B′) design, respectively, with the same $\$_W$ and $\$_L$ levels as above and with $p_W = 2/8, 4/8$ and $p_L = 2/8, 4/8$ in each case.

Procedure

Each student completed four different rating tasks. In each case, instructions and stimuli were presented to the students on a CRT display housed in individual rooms. In Task 1 students were presented with all of the Set A and Set B gambles, one at a time in a random order, and were asked to rate the attractiveness of each on a 100-point scale ranging from "Very Unattractive" to "Very Attractive." Task 2 consisted of having students rate their strength of preference for all of the possible pairs of the 16 gambles in each of Sets A′ and B′. For each pair, the students indicated strength of preference on a 20-point scale ranging from 1 (Strongly Prefer Gamble 1) to 20 (Strongly Prefer Gamble 2). Tasks 3 and 4 were comparable to Tasks 1 and 2 except that in Task 3 "riskiness" was substituted for "attractiveness" and in Task 4 the 20-point scale ranged from "Gamble 1 Is Much More Risky" to "Gamble 2 Is Much More Risky."

Each student attended four separate sessions to complete the tasks, each session being separated by a period of 1 to 3 days. In all cases, three replications of each single stimulus and paired comparison judgment were obtained from the students. For one-half of the students, risk ratings (Tasks 3) were

completed first, followed by the paired comparison difference in risk (Task 4), attractiveness (Task 1), and paired comparison strength of preference judgments (Task 2). For the remaining subjects the order was attractiveness (Task 1), paired comparison strength of preference (Task 2), risk (Task 3), and paired comparison difference in risk judgments (Task 4).

Results

For each student mean attractiveness and riskiness data matrices were formed from the single stimulus ratings for both the Set A and Set B gambles by averaging the three replications of each judgment. These data were then used as input for four separate three-factor within-subjects ANOVAs. A summary of these ANOVA results is presented in Table 1. As expected, the main effects of the $\$_W$, $\$_L$, p_W, and p_L factors were all highly significant ($p < .001$ in each case) in the analyses for both the attractiveness and risk data. In addition, despite the restriction of the levels of the p_W and p_L factors in the design of the Set A and Set B gambles, a number of the interactions among the probabilities and payoffs were significant. Although the relative size of these interaction effects was much smaller than the separate contributions of the payoffs and probabilities in each analysis, these results indicate that, at least across subjects, attractiveness and risk judgments do not follow a simple additive model.

Table 1
Summary of Three-Factor Within Subjects ANOVAs from Attractiveness and Riskiness Data for Set A and Set B Gambles

Set A Source	df	Attractiveness F	Riskiness F
$\$_L$	3, 54	25.99 ***	20.83 ***
p_W	2, 36	71.65 ***	22.59 ***
p_L	3, 54	60.32 ***	64.18 ***
$\$_L \times p_W$	6,108	.41	1.04
$\$_L \times p_L$	9,162	5.40 ***	4.16 ***
$p_W \times p_L$	6,108	2.43 *	.41
$\$_L \times p_W \times p_L$	18,324	1.96 *	1.84 *
Set B Source	**df**	**Attractiveness F**	**Riskiness F**
$\$_W$	3, 54	24.01 ***	15.02 ***
p_W	3, 54	71.68 ***	25.34 ***
p_L	2, 36	53.32 ***	58.31 ***
$\$_W \times p_W$	9,162	6.87 ***	5.01 ***
$\$_W \times p_L$	6,108	8.78 ***	2.21 *
$p_W \times p_L$	6,108	5.32 ***	3.52 **
$\$_W \times p_W \times p_L$	18,324	3.51 ***	4.08 ***

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

MDPREF and NMRG Analyses

Two separate MDPREF analyses were computed on the attractiveness and risk data for the Set A and Set B gambles. In both cases a three-dimensional solution was clearly evident. For the Set A gambles the three-dimensional solution accounted for 90.7% of the variance in the students' data. For the Set B gambles this value was 88.8%. The inclusion of a fourth dimension in each case added little; the fourth dimension accounted for 1.4% and 1.6% of the variance in the data for the two sets of gambles, respectively. This high degree of fit for the three-dimensional solutions was also reflected by the

Figure 2
Dimension 1 (Amount to Lose) and Dimension 3
(Probability of Losing) from the
Rotated MDPREF Solution of Attractiveness and Risk Ratings for
Set A Gambles, with Numbers in Each of the 16 Three-Stimulus Groupings (2,3,4)
Representing the Probabilities of Winning (2/8, 3/8, 4/8) for the Gambles,
Small Arrows Indicating Directions of Subjects' Attractiveness
Vectors and Large Arrows Indicating Directions of Subjects' Risk Vectors.

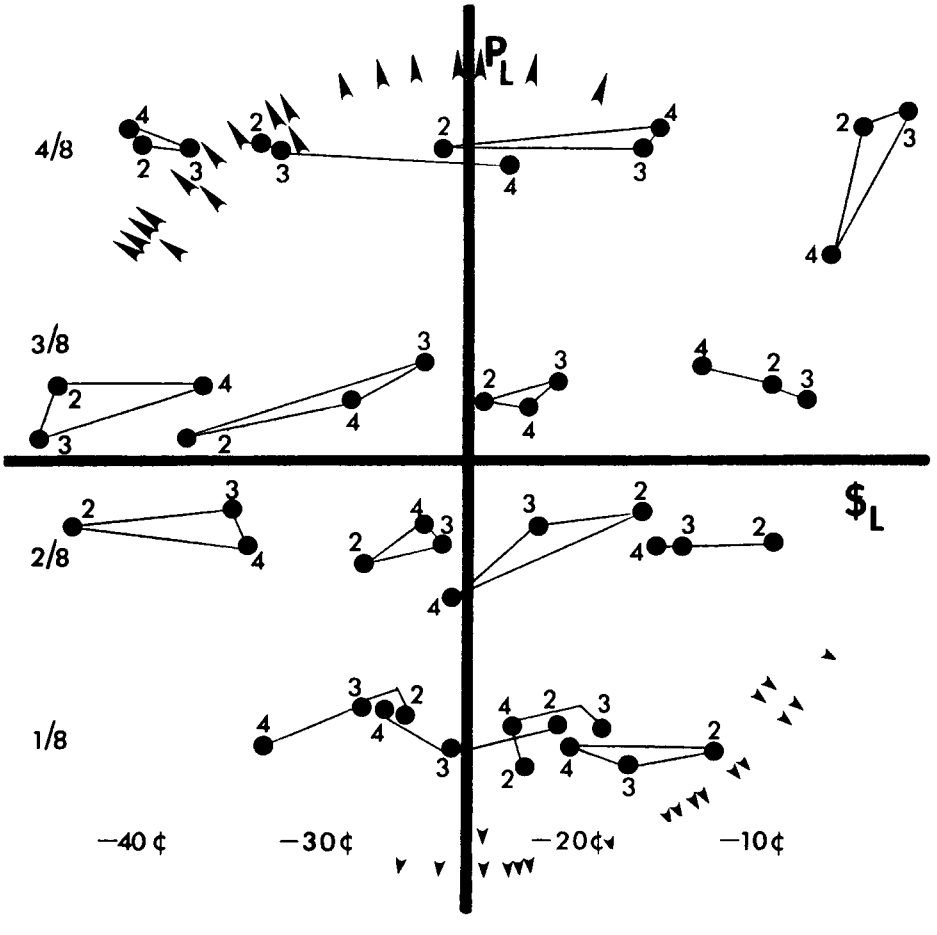
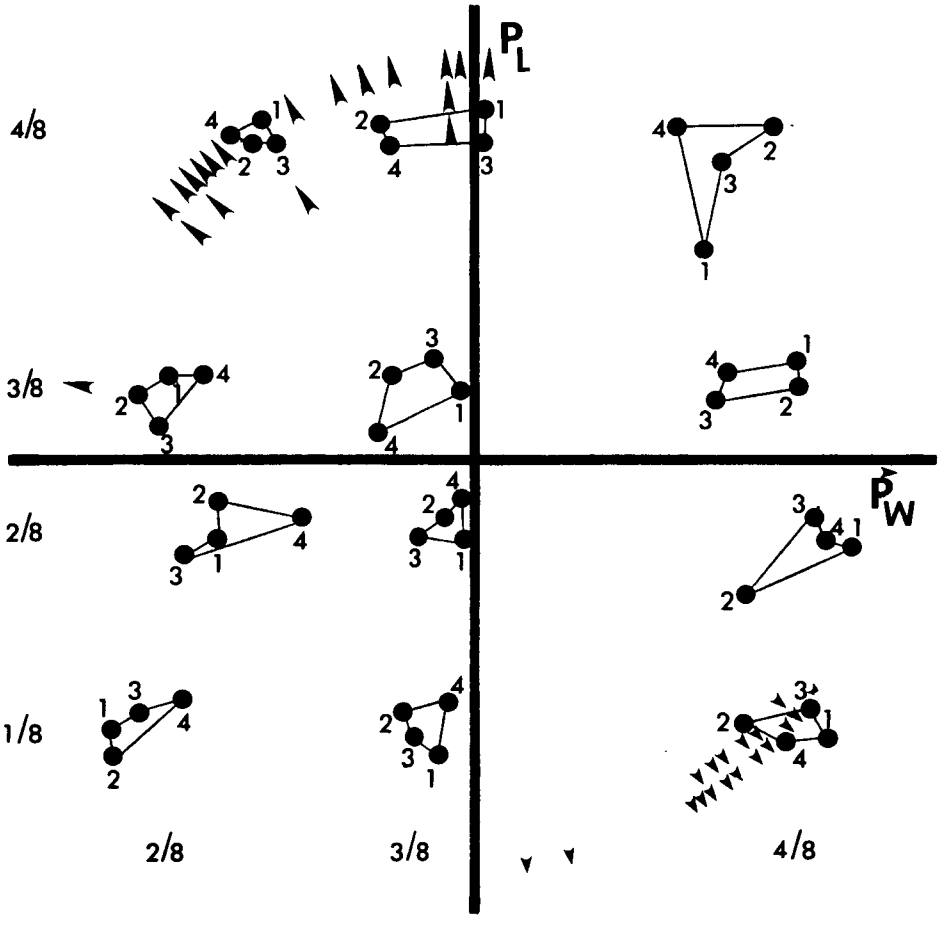


Figure 3
Dimension 2 (Probability of Winning) and Dimension 3 (Probability of Losing) from the Rotated MDPREF Solution of Attractiveness and Risk Ratings for Set A Gambles, with Numbers in Each of the 12 Four-Stimulus Groupings (1,2,3,4) Representing the Amounts to Lose (−10¢, −20¢, −30¢, −40¢) for the Gambles

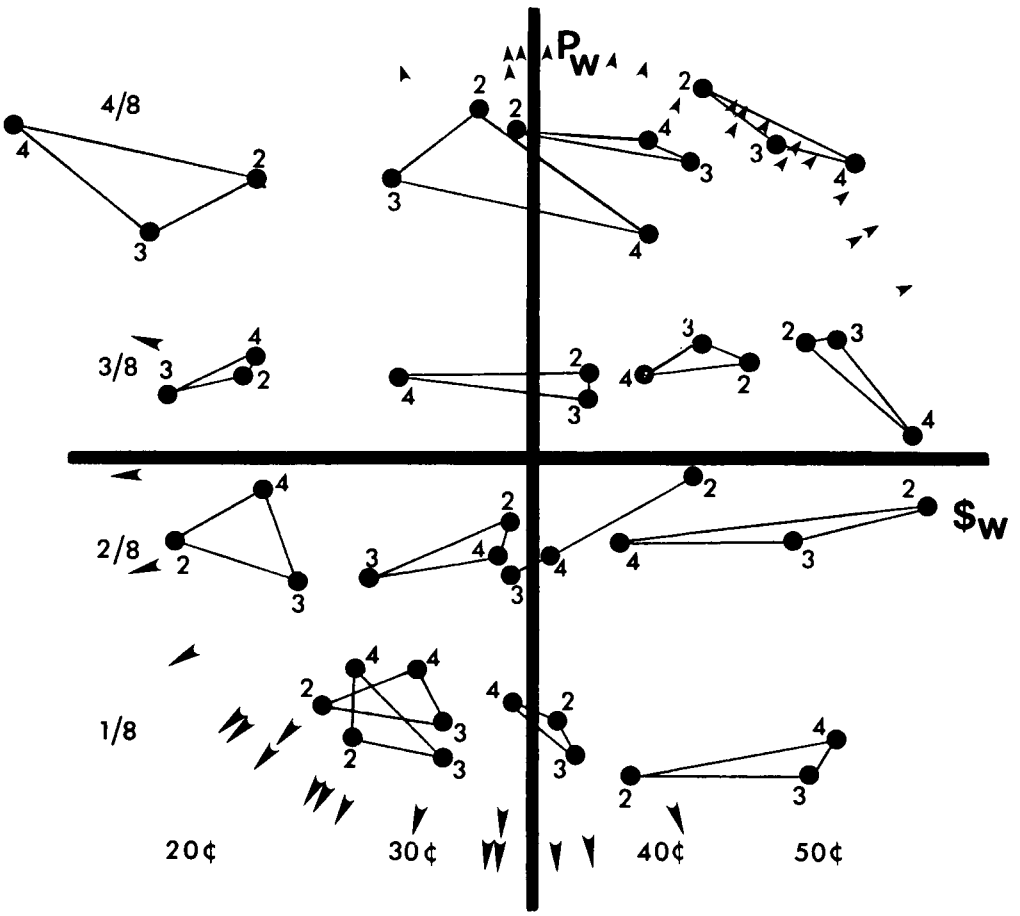


fit of the individual students' data. For the Set A gambles, the root-mean-squared correlations between the students' data and the projections of the gambles on the students' fitted vectors were .936 for both the attractiveness and risk data. For the Set B gambles, these root-mean-squared correlations were .936 and .920, respectively.

Using a generalized rotation procedure outlined in Nygren (1977)[1], the MDPREF solutions for the Set A and Set B gambles were rotated to maximum congruence with two normalized target spaces where the dimensions were $\$_L$, p_W, and p_L, and $\$_W$, p_W, and p_L. Figures 2 through 5 present the representa-

[1]Since the MDPREF model can be stated as $\mathbf{S} = \mathbf{Y} \cdot \mathbf{X}'$, a general transformation $\mathbf{T}$, can be applied to $\mathbf{X}$ to define a new space $\mathbf{X}^* = \mathbf{X} \cdot \mathbf{T}$, and a new set of vector locations $\mathbf{Y}^* = \mathbf{Y} \cdot (\mathbf{T}')^{-1} \cdot \mathbf{X}^*$ and $\mathbf{Y}^*$ form a valid representation of the data, since $\mathbf{Y}^* \cdot \mathbf{X}^{*\prime} = \mathbf{Y} \cdot (\mathbf{T}')^{-1} \cdot \mathbf{T}' \cdot \mathbf{X}' = \mathbf{Y} \cdot \mathbf{X}' = \mathbf{S}$.

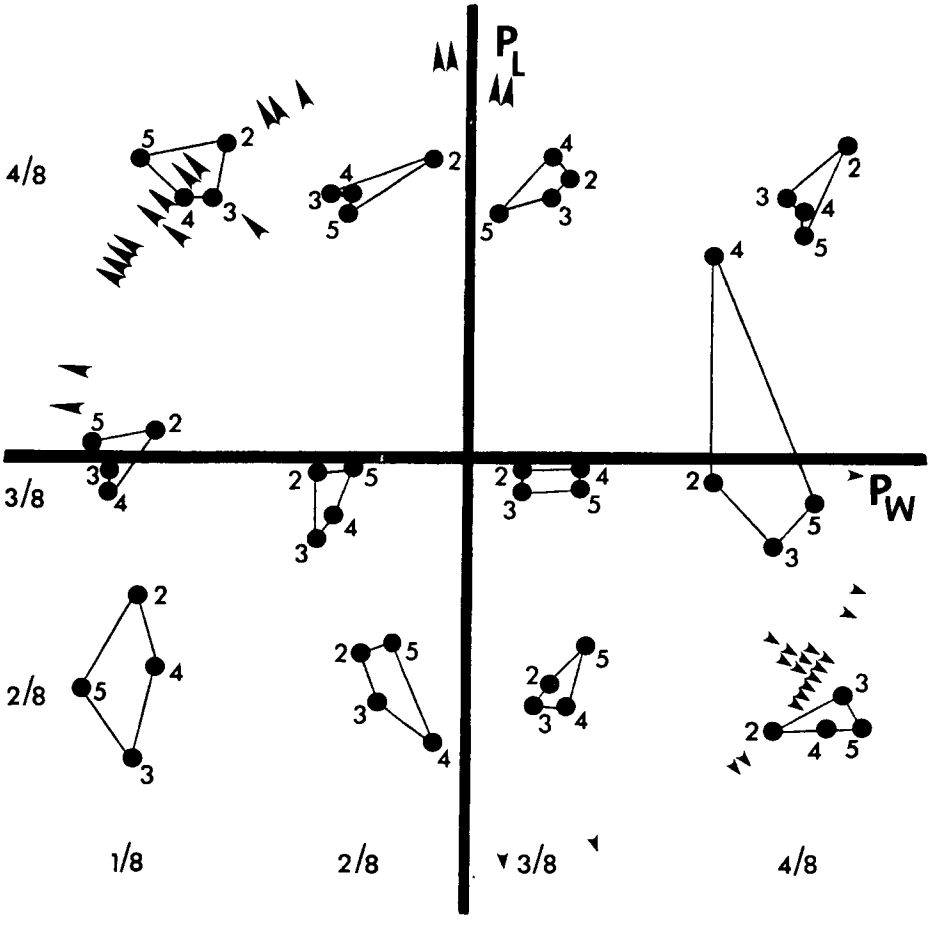
Figure 4
Dimension 1 (Amount to Win) and Dimension 2 (Probability of Winning) from the Rotated MDPREF Solution of Attractiveness and Risk Ratings for Set B Gambles, with Numbers in Each of the 16 Three-Stimulus Groupings (2,3,4) Representing the Probabilities of Losing (2/8, 3/8, 4/8) for the Gambles



tions of the gambles and the students' vectors in these rotated spaces. Although there are some deviations from the target locations of the stimuli, the payoff and probability dimensions are clearly identifiable from the figures. It appears from these MDPREF dimensions that students were using the objective payoff and probability values of the gambles in making their judgments of risk and attractiveness.

Two other implications can be made from these figures. First, the amount of individual differences in the importance of these dimensions in determining risk and attractiveness is extensive. The small arrows in Figures 2 through 5 represent the locations of the students' attractiveness vectors in the spaces; the large arrows indicate the risk vectors. There does not appear to be any obvious pattern

Figure 5
Dimension 2 (Probability of Winning) and Dimension 3 (Probability of Losing) from the Rotated MDPREF Solution of Attractiveness and Risk Ratings for Set B Gambles, with Numbers in Each of the 12 Four-Stimulus Groupings (2,3,4,5) Representing the Amounts to Win (20¢, 30¢, 40¢, 50¢) for the Gambles



as to which dimensions are most important for the students. Rather, each dimension appears to be most important for at least some students.

In addition, the interaction effects found in the ANOVA results are evident in the figures. In particular, Figures 1 and 3 indicate that the objective levels of p_L and $\$_L$, and p_W and $\$_W$ appear to interact in a predictable manner. In Figure 1, for example, it can be seen that as p_L decreases, the perceived differences between levels of $\$_L$ are reduced.

In addition to the MDPREF analyses, a separate NMRG scaling of the 48 gambles in each set was done for the attractiveness and risk data for each student. Recall that for the NMRG procedure the stimulus dimensions must be specified a priori. Because of the high degree of interpretability of the MDPREF dimensions, it was expected that the NMRG solutions across subjects would result in good

fits to the data when the objective payoff and probability dimensions were used. The results support this expectation. Since NMRG is a monotone regression procedure, a Kendall's tau coefficient was computed between the observed data and the predicted scale values for the stimuli. For the Set A gambles, the root-mean-squared tau value across subjects was .761 for the attractiveness data and .743 for the risk data. For the Set B gambles these root-mean-squared tau values were .752 and .710.

Within-Subject Validation

The high degree of fit of the subjects' data to the MDPREF and NMRG scaling solutions are encouraging but offer no direct evidence of the validity of the additive property that is implied by each model. In order to provide a test of additivity, the predicted strength of preference and difference in risk values were computed for each student using Equations 2 and 4 and the derived scale values of the gambles from the MDPREF and NMRG analyses of the Set A and Set B gambles. These predictions were compared with the independently obtained paired comparison judgments from the 16 gambles in Sets A′ and B′.

For each student the proportion of times the predicted paired comparison preference or difference in risk order did not match the order from the corresponding observed judgment was computed. Table 2 presents for each student a summary of this error of prediction data for both models. The error proportions are based on the combined 240 comparisons from the Set A′ and Set B′ gambles (120 comparisons of the 16 gambles in each set). In addition, the average correlation between predicted difference judgments and observed difference judgments were computed for each subject and are also shown in the table.

The results of Table 2 indicate that both MDPREF and NMRG scaling solutions predict the order of the paired comparison difference judgments surprisingly well. For the MDPREF predictions the median proportion of errors across subjects was .100 (24/240) and .104 (25/240) for the preference and risk data, respectively. For the NMRG predictions these median error proportions were .133 (32/240) and .138 (33/240). Within each model there was no difference in predictive ability for preference and risk judgments ($p > .10$ in each case based on Wilcoxon signed rank tests). There was, however, evidence that the MDPREF scale values produced significantly fewer errors in prediction for both the preference data and the risk data than did the NMRG values. A Wilcoxon signed rank test indicated that the median difference in error proportions between models was significantly different from zero in each case (Median difference = $-.029$, $z = 3.56$, $p < .001$ for preference; Median difference = $-.029$, $z = 2.00$, $p < .05$ for risk).

Although these small overall error proportions in prediction appear to support the MDPREF and NMRG models, a more detailed test for additivity was desired. The 120 comparisons for each of the Set A′ and Set B′ gambles were divided into a set of 74 dominance comparisons and 46 tradeoff comparisons. For each subject the proportions of errors of prediction were then computed for each of these latter sets of comparisons. These error proportions are shown in Table 3 for both data types and models. Again, in each case the data from the Set A′ and Set B′ gambles were combined, yielding a total of 148 dominance and 92 tradeoff comparisons for each subject.

The results of Table 3 clearly contrast the dominance and tradeoff comparisons. For both the MDPREF and NMRG predictions, few errors were obtained among dominance comparisons. The median proportions of errors for these dominance pairs were .020 (3/148) and .034 (5/148) for the preference and risk data from MDPREF and .034 (5/148) and .061 (9/148) from NMRG. These low error rates appear to support the hypothesis that the students were, in fact, using the payoff and probability dimensions in making their judgments and that simple independence between pairs of these dimensions holds.

Table 2

Proportion of Errors and Mean Correlations Between Predicted and Observed Paired Comparison Preference and Risk Differences for MDPREF and NMRG Models

	MDPREF				NMRG			
Sub	Preference		Riskiness		Preference		Riskiness	
	P	$\bar{r}$	P	$\bar{r}$	P	$\bar{r}$	P	$\bar{r}$
1	.038[a]	.886	.046	.903	.096	.873	.063	.906
2	.063	.735	.054	.755	.083	.712	.067	.734
3	.079	.808	.100	.825	.108	.785	.096	.814
4	.054	.777	.046	.841	.138	.749	.142	.864
5	.138	.950	.175	.800	.163	.904	.217	.597
6	.108	.731	.163	.807	.125	.708	.163	.707
7	.079	.791	.054	.790	.117	.780	.138	.816
8	.188	.659	.196	.146	.196	.612	.154	.164
9	.058	.726	.025	.884	.050	.550	.054	.860
10	.075	.852	.100	.887	.100	.819	.083	.793
11	.071	.913	.142	.937	.058	.878	.129	.941
12	.150	.839	.054	.971	.250	.814	.133	.956
13	.100	.828	.104	.931	.133	.772	.058	.889
14	.096	.863	.125	.805	.104	.875	.121	.827
15	.208	.927	.204	.859	.283	.913	.275	.853
16	.117	.769	.250	.916	.196	.761	.217	.911
17	.208	.775	.442	.699	.254	.736	.542	.649
18	.183	.813	.096	.937	.208	.674	.200	.865
19	.167	.743	.225	.683	.313	.674	.338	.640
Med	.100		.104		.133		.138	
RMS		.813		.828		.774		.798

[a]Each error proportion is based on 240 comparisons, 120 from each of the Set A and Set B gambles.

The error proportions for the tradeoff comparisons, however, were much higher for nearly all subjects for both models and both data types. In fact, for several subjects, these tradeoff error proportions were not significantly smaller than the chance level of .50 (critical value of $\hat{p} = .379$ for $\alpha = .01$). The median error proportions for these tradeoff comparisons are also shown at the bottom of Table 3. In all four cases Wilcoxon signed rank tests indicated that the median difference between corresponding dominance and tradeoff error proportions was significantly smaller than zero across subjects ($p < .01$ in each case, cf. bottom of Table 3).

This consistent difference between dominance and tradeoff error proportions provides support for the hypothesis that simple independence holds in the preference and risk data, but double cancellation and/or joint independence between all pairs of factors do not. It is important to recognize, however, that in practice even if additivity holds in a person's data, P_T and P_D might be expected to be greater than zero because of imperfect discriminability and P_T might be expected to be greater than P_D because tradeoff comparisons may be inherently more difficult to make than are dominance com-

Table 3

Proportion of Errors for Dominance and Tradeoff Paired Comparisons Based on MDPREF and NMRG Predictions

Sub	MDPREF				NMRG			
	Preference		Riskiness		Preference		Riskiness	
	P_D	P_T	P_D	P_T	P_D	P_T	P_D	P_T
1	.000[a]	.098[b]	.007	.109	.000	.250	.007	.152
2	.000	.163	.000	.141	.000	.217	.054	.087
3	.000	.207	.074	.141	.000	.283	.047	.174
4	.000	.141	.020	.087	.027	.315	.020	.337
5	.000	.359	.020	.424	.000	.424	.047	.489
6	.007	.272	.020	.391	.007	.315	.020	.391
7	.007	.196	.034	.087	.027	.261	.162	.098
8	.014	.467	.034	.457	.041	.446	.027	.359
9	.020	.120	.007	.054	.034	.076	.027	.098
10	.020	.163	.020	.228	.027	.217	.068	.120
11	.020	.152	.068	.261	.047	.076	.068	.228
12	.020	.359	.061	.043	.135	.435	.149	.109
13	.034	.207	.034	.217	.027	.304	.000	.152
14	.041	.185	.081	.196	.054	.185	.081	.185
15	.068	.435	.230	.163	.189	.435	.236	.337
16	.095	.152	.318	.141	.236	.130	.270	.130
17	.108	.370	.412	.489	.196	.348	.547	.533
18	.169	.207	.088	.109	.155	.293	.223	.163
19	.223	.076	.223	.228	.385	.196	.284	.424
Med	.020	.196	.034	.163	.034	.283	.061	.174
Med Diff	-.163		-.077		-.234		-.104	
r	-.127		.211		-.037		.395	

[a]Error proportions for dominance tests are based on 148 comparisons, 74 from each of Sets A and B.

[b]Error proportions for tradeoff tests are based on 92 comparisons, 46 from each of Sets A and B.

parisons. This last point suggests a potential confounding in using difference estimates $P_T - P_D$ in a test of additivity. Admittedly, a completely satisfactory elimination of this potential confounding is not possible without an explicit error theory for P_D and P_T. There are several results, however, that suggest that the difference between dominance and tradeoff error proportions is a function of nonadditivity rather than merely level of difficulty.

The differences between P_D and P_T values in Table 3 are, for many subjects, very large. It is difficult to imagine that a difference in the level of difficulty in making dominance versus tradeoff comparisons could be responsible for P_D values for some subjects being virtually zero (error-free discrimination) and the corresponding P_T values being at or near the chance level (random discrimination). In fact, even if one-half of all tradeoff errors were attributed to being due to level of difficulty

and they were removed from Table 3 (i.e., cut P_T by one-half in each case), for the 65 cases in Table 3 where P_T was initially greater than P_D, in 52 (or 80.0% of them) one-half of P_T is still greater than P_D. Further, there is no evidence to suggest that the differences between dominance and tradeoff error proportions were due merely to unreliability or to error variance in the subjects' data. If such were the case, one would expect a high positive correlation across subjects between the dominance and tradeoff error proportions. This was not found. The bottom row of Table 3 shows the obtained correlations between these error proportions in each of the four model by data type combinations. None were significant ($p > .05$ in each case).

Discussion and Conclusions

Two sets of 48 three-outcome gambles were constructed from a three-factor design in which the factors were either $\$_L$, p_W, and p_L or $\$_W$, p_W, and p_L. Previous research (Payne, 1975; Shanteau, 1974, 1975; Slovic & Lichtenstein, 1968) has suggested that individuals use the explicit payoffs and probabilities in making judgments of the attractiveness and riskiness of gambles. The MDPREF solutions for the Set A and Set B gambles, as illustrated in Figures 2 through 5, support this hypothesis. In each figure the constructed factors are clearly definable, although some perturbations among the stimuli exist.

It is argued here, however, that the interpretability of the derived dimensions, in conjunction with the high degree of fit of the subjects' data to the model, cannot and should not be interpreted as sufficient criteria to validate the underlying additive model. The stimuli used in this study were chosen specifically to illustrate this point. It was expected that payoffs and probabilities would be combined in a nonadditive fashion to produce the subjects' judgments. The ANOVA results in Table 1 demonstrated the expected interaction effects among the factors.

The overall error proportions of prediction, as shown in Table 2, however, were small across subjects despite this interaction. This is a particularly interesting result in light of the fact that individual differences in the importance of the payoff and probability dimensions for both risk and attractiveness judgments are extensive. Although the subjects' vectors in Figures 2 through 5 indicate, as would be expected, that perceived risk increases and perceived attractiveness decreases as Amount to Lose ($\$_L$) and Probability of Losing (p_L) increase and as Amount to Win ($\$_W$) and Probability of Winning (p_W) decrease, it is clear that no single payoff or probability dimension was dominant across subjects for either type of judgment.

The error proportions shown in Table 3 illustrate that the predictive success of the MDPREF and NMRG models to the independently obtained paired comparison data is misleading with respect to additivity. There is a consistently significant discrepancy between predictive success for the dominance pairs and the tradeoff pairs. The preference and risk orders for the dominance pairs were predicted extremely well—in some cases, without any errors. The predicted orders for the tradeoff pairs were much worse—in some cases, the error proportions were not significantly better than would be expected by chance.

These results have several important implications for researchers using MDPREF, NMRG, or other additive conjoint scaling techniques. If, as appears to be the case in this study, simple independence among the factors holds in the data, the goodness of fit of the model to the data may be misleading. A moderate to high degree of fit will be exhibited in the scaling solutions even if additivity is violated. It appears that these conjoint scaling techniques can adequately reflect some aspects of the processing model used by the subjects—namely, the dimensional structure of the stimuli and individual differences estimates. Yet, in a predictive sense these additive models are unacceptable. For nearly all of the subjects in this study, the error proportions for tradeoff comparisons were high. These

tradeoff comparisons are, however, the very comparisons of interest in applied decision-making research. It is important to recall in this connection that the nonadditivity effect was purposefully minimized in this study by restricting the range of the p_W and p_L values. Had a broader range of values for these two factors been used, it is likely that the tradeoff predictions would have been even worse.

Why these scaling models do so well in a goodness-of-fit sense, even though additivity is violated, appears to be a function of the stimulus design. For small full factorial designs, which are typically used in conjoint scaling research, the ratio of the number of implicit dominance pairs to the number of tradeoff pairs will usually be greater than 1.0. Hence, within the scaling models, if there is consistently little error in the subjects' data as it reflects the order of dominance pairs, it seems plausible that an acceptable additive scaling solution (as inferred from a traditional correlational goodness-of-fit measure) is almost guaranteed. Even for as large a design as the $4 \times 4 \times 3$ design used in this study, the ratio of dominance to tradeoff pairs is only slightly less than 1.0 (553/575 = .972).

The value of conjoint scaling models as data reduction techniques has been clearly demonstrated in a number of research areas. Yet it is argued here that if one is interested in using an additive conjoint scaling model to predict individual tradeoff comparisons among multifactor choice alternatives, then an assessment of the additivity assumptions seems imperative. A validation procedure as illustrated here can be used to evaluate the predictive success of the model for the critical tradeoff pairs. In addition, it would seem worthwhile to reconstruct the traditional goodness-of-fit measures used in these scaling procedures to reflect predictive success for dominance and tradeoff comparisons separately. In this way, additivity could be more meaningfully evaluated.

References

Carroll, J. D. *Categorical conjoint measurement.* Paper presented at the Mathematical Psychology Meeting, Ann Arbor, MI, 1969.

Carroll, J. D. Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 1). New York: Seminar Press, 1972.

Coombs, C. H., & Bowen, J. N. Additivity of risk in portfolios. *Journal of Experimental Psychology,* 1970, *10,* 43-46.

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin,* 1974, *81,* 95-106.

Eckart, C., & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika,* 1936, *1,* 211-218.

Green, P. E., & Rao, V. R. Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research,* 1971, *8,* 355-363.

Green, P. E., & Srinivasan, V. Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research,* 1978, *5,* 103-123.

Holt, J. O., & Wallsten, T. S. *CONJOINT: A PL/1 program for evaluating conjoint measurement axioms* (Technical Report). L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC, 1974.

Johnson, R. M. A simple method for pairwise monotone regression. *Psychometrika,* 1975, *40,* 163-168.

Krantz, D. H., & Tversky, A. Conjoint-measurement analysis of composition rules in psychology. *Psychological Review,* 1971, *78,* 151-169.

Kruskal, J. B. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society,* 1965, *27,* 251-263.

Nygren, T. E. The relationship between the perceived risk and attractiveness of gambles: A multidimensional analysis. *Applied Psychological Measurement,* 1977, *1,* 565-579.

Payne, J. W. Alternative approaches to decision making under risk: Moments versus risk dimensions. *Psychological Bulletin,* 1973, *80,* 439-453.

Payne, J. W. Relation of perceived risk to preferences among gambles. *Journal of Experimental Psychology,* 1975, *104,* 86-94.

Shanteau, J. C. Component processes in risky decision making. *Journal of Experimental Psychology,* 1974, *103,* 680-691.

Shanteau, J. C. An information integration analysis of risky decision making. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes.* New York: Academic Press, 1975.

Slovic, P., & Lichtenstein, S. Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology Monograph*, 1968, *78* (3, part 2).

Srinivasan, V., & Shocker, A. D. Linear programming techniques for multidimensional analysis of preferences. *Psychometrika*, 1973, *38*, 337-369. (a)

Srinivasan, V., & Shocker, A. D. Estimating the weights for multiple attributes in a composite criterion using pairwise judgments. *Psychometrika*, 1973, *38*, 473-493. (b)

Acknowledgment

This research was supported in part by a University Small Research Grant from The Ohio State University.

Author's Address

Send requests for reprints or further information to Thomas E. Nygren, Department of Psychology, Ohio State University, 404C W. 17th Avenue, Columbus, OH 43210.