Research Article

DOES EVIDENCE PRESENTATION FORMAT AFFECT JUDGMENT? An Experimental Evaluation of Displays of Data for Judgments

Alan Sanfey and Reid Hastie
University of Colorado at Boulder

Abstract—*Information relevant to a prediction was presented in one of eight formats: a table of numbers, a brief text, a longer biographical story, and five different types of bar graphs. Experimental participants made judgments of marathon finishing times based on information about the runners' ages, prior performance, training, and motivation. A regression analysis was used to assess the individual judges' relative weighting of the various types of information relevant to their predictions. The different formats for displaying information yielded different levels of accuracy and patterns of information utilization. In accordance with an explanation-based decision model, the text and story displays induced the heaviest reliance on information about runners' motivation and prior performance and produced the most accurate judgments of marathon finishing times.*

People make dozens of quantitative judgments every day: How much will the groceries cost? What will the temperature be tomorrow? What will the interest rate be on the mortgage in a year? The information relevant to these judgments comes in many formats, including conversations, text, tables, and graphs. When people are serious about making judgments carefully, modern computer software provides a great variety of possible formats in which to present quantitative information. The freedom to choose formats raises the practical question of which types of displays produce the best judgments.

Only a few empirical studies have attempted to assess the efficacy of alternate formats for displaying evidence in decision-making tasks. The basic experimental paradigm is obvious: select a fixed set of judgment-relevant data, display the data in various formats, and then measure the accuracy and hypothesized mediators of the judgment. One exemplary study was conducted by MacGregor and Slovic (1986), who manipulated the format in which their subjects received information used to make estimates of the finishing times for male runners in a local marathon. The participants in the study were given four informational cues for each runner on which to base their estimates: the runner's age, the number of training miles run, the runner's fastest time for a 10-km run, and a self-rating by the runner of how motivated he was to run a fast race. These cues were displayed in four different graphical formats: a bar graph display, a deviation display, a spoke display, and a face display. Initially, the face display (Chernoff, 1973) produced more accurate judgments than the other displays; however, the most valid cues were represented by the most salient facial features, possibly confounding the results. Indeed, when the face display was tested again with a haphazard assignment of cues to facial features, it did not prove substantially more effective than the other displays.

Two principles to predict which information will tend to dominate judgments emerged from MacGregor and Slovic's findings and from other empirical studies. First, displays vary in terms of the degree of salience they impart to their components. Information that is perceptually distinctive in a given display tends to have a heightened impact on any inference or decision based on that display. Second, commensurability effects are usually present: The more similar (compatible) the format of a stimulus and a response dimension, the greater the impact of that dimension on the responses (Hastie, Hammerle, Kerwin, Croner, & Herrmann, 1996; Slovic, Griffin, & Tversky, 1990). For example, when valuation responses are required on a dollar metric scale, stimulus information expressed in dollars has a greater impact than information expressed in other metrics.

We also propose that general characteristics of the format in which information is presented influence how information is weighted in judgment tasks. For example, information about motivation or emotional state, or perhaps about biographical events, might receive most consideration and be weighted most heavily in a format that is consistent with narrative communication. In contrast, information that is usually communicated numerically may receive greatest attention and weight when it is presented in a tabular format.

The weight accorded to various types of information is important because it determines accuracy (Hastie & Rasinski, 1988). One of the most popular methods to decompose accuracy into its components is to experimentally assess the behavioral cue utilization weights and, under conditions in which the environment is also well understood, to compare them with the environmentally adaptive cue validity weights (Brunswik, 1956). The highest levels of accuracy will occur when the cues in the environment are utilized in a manner that reflects their true validity. Thus, the most effective displays will be those in which the cue utilization weights most closely match the environmentally valid cue-criterion coefficients (this seems to have been what occurred inadvertently when MacGregor & Slovic, 1986, created their first set of Chernoff face stimuli). When display factors like salience and compatibility make the most ecologically valid cues the most prominent to the judge, accuracy will be maximized.

The present experiment extends MacGregor and Slovic's exploration of the effects of display formats in four directions. First, bar graphs are probably the most popular graphical format utilized in decision-making tasks, so we added several variations on the basic bar graph display. MacGregor and Slovic utilized graphs that displayed four vertical bars, each one corresponding to the value of one of the cues. However, the graph did not contain the numerical cue values, so the judge had to estimate precise values if he or she thought that numerical information was more useful. Therefore, we included a condition that presented the bars as in MacGregor and Slovic's study, but with the addition of the numerical cue values. Furthermore, in this particular study, the four cues were not related to the response dimension in a consistent, direct fashion. For two of the cues, total miles trained and the self-rating of motivation, higher values predicted faster times, but for the other two, runner's age and fastest 10-km time, lower values predicted faster times. Applying the compatibility principle, we included two more bar graph conditions in which the heights of the

Address correspondence to Alan Sanfey, Psychology Department - CB 345, University of Colorado at Boulder, Boulder, CO 80309; e-mail: asanfey@psych.colorado.edu.

bars were all consistently related to finishing time, one in which taller bars indicated faster times (i.e., we displayed the inverses of age and fastest 10-km time) and one in which taller bars indicated slower times (i.e., we displayed the inverses of miles trained and motivation). We predicted that a consistent display would lead to more accurate judgments than the mixed display.

Second, because numerical data are most often displayed in a table format (Jarvenpaa & Dickson, 1988), we included a table display in our experiment. A simple prediction, from our speculations about format effects, was that nonnumerical conceptual cues, like the runner's motivation, would receive distinctively low weights when presented as tabled numbers.

Third, data relevant to decisions is often conveyed in text or spoken communications. Recent research has indicated that when the relevant propositional evidence concerns motivated human behavior (e.g., committing a crime, competing in a sports event), people make their judgments based on summaries in the form of narrative stories constructed to explain the evidence (Pennington & Hastie, 1993). Evidence for this theory comes primarily from the legal domain, in which jurors usually construct summary mental explanations of the trial evidence before making judgments of guilt or innocence. If it is the case that judgments of motivated human behavior are based on narrative causal explanations, we would expect that providing the cue values in such a format would facilitate the judgment process, as it would ease the construction of explanatory narratives from evidence cues. We would also expect that cues that represent motivations would receive greatest weight when evidence is displayed in a text format, as these cues are highly salient in narrative text representations (cf. Singer & Halldorson, 1996).

Finally, it is of interest to examine whether one can improve judgment by informing judges of how the cues are actually related to the to-be-predicted quantities. We therefore examined a final condition in which participants were told how to weight the four cues optimally; in this condition, information about the runners was displayed in the mixed-graph format.

Our study is a systematic replication of MacGregor and Slovic's method, with our participants judging marathon finishing times based on the same four cues (runners' age, motivation, past race time, and training). We constructed eight evidence displays: a tabular display, a brief text display, a longer story display, and five types of bar graph displays. We hypothesized that the text-based displays would facilitate use of the most compatible cues and would also lead judges to construct a narrative explanation of the data. Because biographical (e.g., past 10-km time) and motivational information are likely to be more prominent in narrative representations, and past research showed these factors to be usually underweighted, we expected that the text-based displays would promote more accurate judgments. Also, we expected the consistent-graph display to promote better judgments than MacGregor and Slovic's mixed-graph display.

METHOD

Participants

The participants in this study were 128 psychology undergraduates (70 females and 58 males) from the University of Colorado, participating for course credit. Prior knowledgeability about marathon running varied, but this individual difference is ignored in our analyses as it was not associated with any of the measures of judgment performance.

Table 1. *Matrix of cue intercorrelations for the study sample of marathon runners*

	Age	Total training miles	Fastest 10-km race	Motivation
Age	—	–.12	.36	–.06
Total training miles		—	–.36	–.26
Fastest 10-km race			—	.15
Motivation				—

Materials

Participants were asked to estimate the time it took each of 35 runners to complete a marathon. Each runner was described in terms of four information cues: age, number of training miles run, fastest 10-km-race time, and a self-rating of the runner's motivation to run the marathon in a fast time. Table 1 displays the set of cue intercorrelations.

These cues were presented in the same order for every runner. In addition, subjects were given an overview of the nature of the marathon; the times of the slowest, fastest, and average runners in the race; and the high and low values of each of the information cues.

Each participant received the cue information in one of the eight experimental display formats. All formats presented the 35 runners in the same order.

- *Table.* The information was represented in the form of a table containing the cue values (see Fig. 1).
- *Text.* The information was represented in the form of a short paragraph containing the relevant information:

  Runner number 5 was 35 years old when he competed in the Trail's End marathon. In the two months prior to the race he ran a total of 368 training miles. His fastest time for a 10km run in the past year was 40 minutes. On the day of the race, he rated himself as moderately motivated to achieve a time goal in this event.

- *Story.* The information was represented in the form of a biographical sketch of each runner containing the relevant information:

  Chuck Norvil was 35 years old at the time of the Trail's End marathon. He works as a dentist in Portland. He was born in Boise, Idaho and went to dental

Age	35
Total Miles	368
Fastest 10km	40
Time Motivation	2

Fig. 1. The table display for the fifth runner.

school at the University of Wisconsin in Madison. He has been married for 8 years, but as yet has no children. As part of his marathon preparation, he ran a total of 368 training miles in the two months prior to the race. His 10km personal best time was 40 minutes. Chuck also enjoys chess and plays golf regularly. On the day of the race, he rated himself as moderately motivated to achieve a time goal in this event.

- *Mixed graph.* The information was represented in the form of a bar graph display, with each absolute cue value (e.g., age, time of fastest 10-km run) denoted by the height of a shaded bar. These values were scaled between 0 and 1, where 0 was the lowest value of that cue and 1 was the highest. (See Fig. 2.)
- *Mixed graph with numbers.* The information was represented as in the mixed-graph format, with the addition of the actual cue values, each presented under the appropriate bar. (See Fig. 3.)
- *Mixed graph with weights.* The information was represented as in the mixed-graph format, with the addition of a page of instructions informing the participants as to the optimal weighting of the cues.
- *Consistent-direct graph.* The information was represented in the form of a bar graph display, with each cue value denoted by the height of a shaded bar and the height of the bar indicating how that cue is related to a faster time. The cue values were scaled between 0 and 1, and lower values indicated that a slow time would be expected. (See Fig. 4.)
- *Consistent-inverse graph.* The information was represented in the form of a bar graph display, with each cue value denoted by the height of a shaded bar and the height of the bar indicating how that cue is related to a slower time. The cue values were scaled between 0 and 1, and lower values indicated that a fast time would be expected. (See Fig. 5.)

Procedure

Each participant was randomly assigned to one of the eight conditions and received a booklet containing the assigned display format. Participants received instruction on interpreting the data in the particular format seen. They were then asked to judge the 35 runners' finishing times in hours and minutes. Following this, they answered questions regarding their knowledge of running and marathons. Finally, they briefly described how they made their judgments. The task was self-paced, with no time constraints placed on the subjects, and took approximately 35 min to complete.

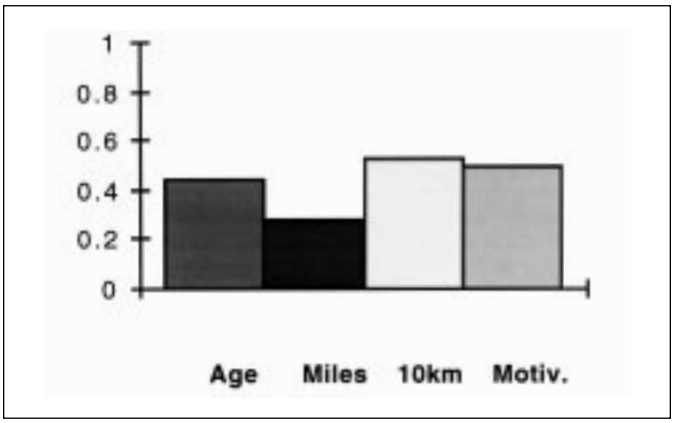

Fig. 2. The mixed display for the fifth runner.

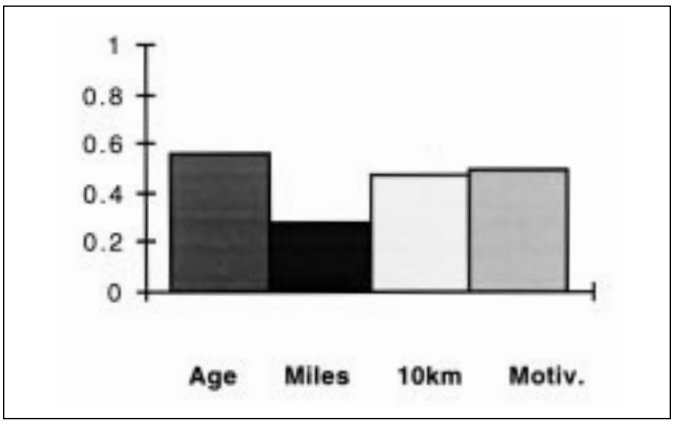

Fig. 4. The consistent-direct display for the fifth runner.

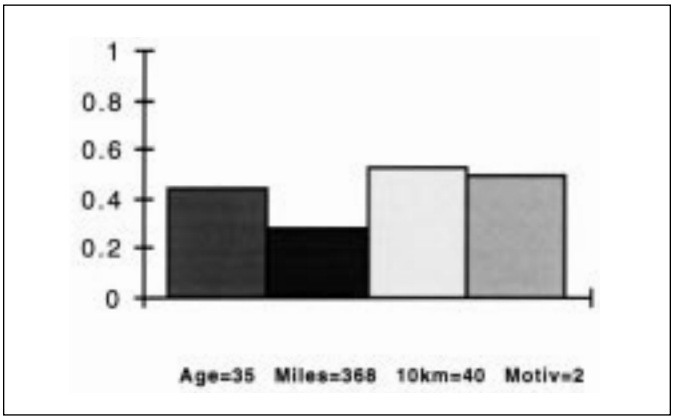

Fig. 3. The mixed display with numbers for the fifth runner.

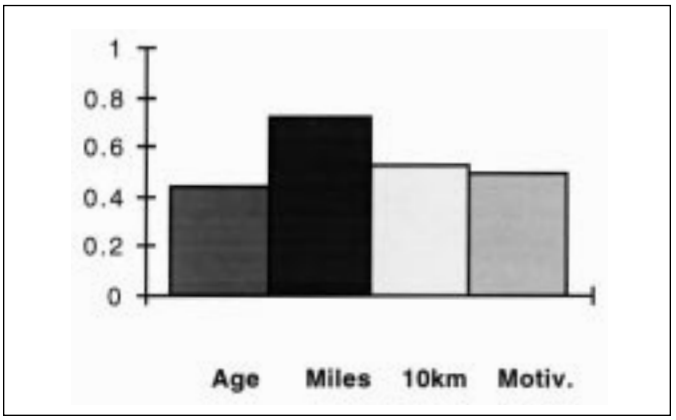

Fig. 5. The consistent-inverse display for the fifth runner.

RESULTS

Accuracy

An accuracy measure, expressed as the mean correlation between the estimated marathon times and the true marathon times, was calculated for each participant. Summary statistics for each display are presented in Table 2. Statistically optimal usage of the four cues would yield an accuracy index of .79. The *r* correlation values were transformed into Fisher *z* scores for the purposes of statistical analysis. The mean accuracy values across the display conditions differed significantly, $F(7, 120) = 3.13$, $p = .001$.

A series of a priori planned comparisons showed that the textual displays produced significantly higher levels of accuracy than the other displays, $F(1, 126) = 15.22$, $p < .001$. Further, the text and story displays produced significantly more accuracy than the bar graph displays, $F(1, 110) = 18.63$, $p < .001$, although the text and story displays did not differ significantly from each other, $F(1, 30) = 0.49$, $p > .05$.

Overall, the mixed-graph displays did not differ significantly in accuracy from the consistent-graph displays, $F(1, 78) = 0.81$, $p > .05$. Likewise, the two consistent bar graphs, the direct and the inverse orientations, did not differ significantly, $F(1, 30) = 0.51$, $p > .05$. Furthermore, the mixed bar graph containing the actual cue values did not differ significantly from the basic bar graph display, $F(1, 30) = 0.01$, $p > .05$, nor did the mixed display containing optimal-weighting instructions differ significantly from the basic mixed graph, $F(1, 30) = 0.19$, $p > .05$.

Relative Cue Weightings

Regression analysis was used to construct a model of each participant's "judgment policy," with standardized beta coefficients representing the impact of each cue on the individual's judgments of race times across the cases judged (Brunswik, 1956; Cooksey, 1996; Hammond, 1955; Stewart, 1988). An analysis of variance was then used to identify any differences in cue weightings across the various display formats. Five planned comparisons between groups were examined. These comparisons examined potential differences between (a) text and story versus other displays, (b) consistent-graph versus mixed-graph displays, (c) consistent-direct-graph versus consistent-inverse-graph displays, (d) mixed-graph-with-instructions versus mixed-graph displays, and (e) text and story versus table displays.

Cue weights indicate the relative strength of correlation between the participants' judgments and each of the information cues (Table 2). On average, the cue of fastest 10-km time was used most heavily, followed by age and total training miles, which were approximately equal in their weighting, and then motivation, which was utilized much less. However, there was a wide variation of cue utilization across display types. Optimal weights, derived from a statistical regression of actual finishing times on the cues are also presented in Table 2. Interestingly, the text displays show the closest match between optimal and actual weights.

Age

Age was utilized significantly less in the text and story than in the other displays, $F(1, 78) = 6.12$, $p = .016$. The difference between the mixed graph with weights and the bare mixed graph was significant, $F(1, 30) = 10.73$, $p = .001$. Participants in the consistent-direct condition utilized this cue significantly more than those in the consistent-inverse condition, $F(1, 30) = 5.63$, $p = .02$.

Miles

This cue was used significantly more in the text and story displays than in the other displays, $F(1, 126) = 3.99$, $p = .042$, and also more in the mixed-graph-with-weights display than in the mixed-graph condition, $F(1, 30) = 8.61$, $p = .004$.

10-km time

This cue was used significantly more in the text and story displays than in the other displays, $F(1, 126) = 8.48$, $p = .004$.

Motivation

Motivation was utilized significantly more in the text and story displays than in the table display, $F(1, 46) = 4.38$, $p = .038$.

Table 2. *Statistics summarizing performance in the judgment task*

Display	Accuracy index[a]	*MSE*	Consistency index[a]	Weight			
				Age	Total training miles	Fastest 10-km race	Motivation
Table	.52 (.05)	719.34	.68 (.05)	.32	–.24	.39	–.08
Text	.60 (.03)	510.14	.69 (.02)	.20	–.33	.53	–.21
Story	.57 (.04)	630.04	.61 (.03)	.26	–.33	.41	–.20
Mixed graph (bare)	.44 (.05)	2,425.62	.52 (.06)	.34	–.15	.33	–.11
Mixed graph (numbers)	.45 (.05)	2,596.87	.71 (.05)	.24	–.29	.30	–.12
Mixed graph (weights)	.42 (.05)	741.28	.68 (.03)	.11	–.34	.38	–.22
Consistent graph (direct)	.44 (.03)	5,471.31	.59 (.04)	.38	–.30	.23	–.18
Consistent graph (inverse)	.37 (.05)	776.80	.67 (.05)	.21	–.18	.33	–.26
Mean	.48 (.04)	1,733.93	.64 (.04)	.26	–.27	.36	–.17
Optimal	.79	233.62	1.00	.18	–.27	.52	–.15

[a]Standard errors are shown in parentheses.

Consistency

The consistency index measures the degree to which an individual participant's judgments are predictable from a linear model based on his or her cue weightings. Statistics in Table 2 indicate how effective each display condition was at promoting consistency. There was no significant effect for the omnibus test, $F(7, 120) = 2.01$, $p > .05$.

DISCUSSION

The results of this study demonstrate that the format in which information is displayed has an impact on how the information is utilized when making judgments. Different presentation formats produced different levels of accuracy, and these differences in accuracy can be understood with reference to different patterns of predictive cue utilization.

Participants who received the information in textual form, whether it was the short description or the longer biographical story, proved to be reliably more accurate in their judgments than those who received the information in graphic or tabular form. We attribute the superiority of the textual displays to their tendency to induce a salience ordering on the cues that reflects environmental cue validities more directly than does the ordering the other formats tend to induce. Specifically, cue utilization measures showed that the text formats led subjects to exhibit cue utilization orderings consistent with the cues' validity coefficients. We believe that the text displays induced an explanation-based judgment strategy that led subjects to utilize information about the runners' motivation and best 10-km-race times more effectively than did subjects in the other format conditions. Subjects' comments on their strategies and some informal think-aloud report trials provided illustrations of the nature of these mediating explanations. In many cases, subjects created impressions of the runners' characters in which motivation served as a prominent organizing principle. Other subjects created biographies of the runners or narrative stories of the race events that, again, thematically emphasized motivation and recent performance information.

There are alternative explanations for the superiority of textual displays, some in terms of explanation-based principles (e.g., all subjects strive to create explanations of the evidence, but this process is simply easier and less error prone in the text format conditions), some in terms of participants' motivation to perform the experimental task (it might be that text displays elicit higher levels of motivation to attend and comprehend the evidence), and others in terms of precision (only the table, text displays, and one of the bar graphs presented exact numerical values of the data). However, we favor our interpretation in terms of an explanation-based judgment strategy.

Another interesting finding concerning accuracy is that there were no significant differences across all forms of the bar graph. Probably the best summary statement is that bar graphs encourage subjects to weight each of the cues approximately equally, at least in comparison with the other formats, but do not induce an advantageous differential weighting of the cues, at least in this particular judgment task. It appears that participants could utilize the consistent and mixed bar graphs equally well, and the addition of exact numerical cue values was not of significant help in making the judgments. This unexpected finding may have been due to the experimental procedure, which included detailed instructions. This instruction may have been effective enough to eliminate initial differences in comprehensibility for the bar graph formats, although the practice given was minimal.

In one condition, we provided bar graph displays plus written instructions on how to weight the cues to maximize accuracy. However, as is common in multicue judgment tasks, the extra instructions did not improve accuracy. This result should be taken as a reminder that judgment skills cannot be magically induced by simple instructional interventions. Current practice suggests that a combination of experience at the task plus feedback on correct weights and feedback on the judge's own cue utilization is necessary to produce substantial improvements in accuracy (cf. Balzer, Doherty, & O'Connor, 1989).

Conventional wisdom has it that graphic displays make it easier for a judge to assimilate information and make a judgment than does a textual format, but the results of the present study provide some evidence countering this view. In fact, both textual formats produced more accurate judgments than did any of the bar graph displays. These results can be understood by noting the effects of format on cue utilization weights. Some formats (text and story formats) had a tendency to induce weighting patterns that were most consistent with the optimal cue validity weights prescribed by the actual environment of marathon race performances. When the display format and the structure of the environment are consistent, the format will produce superior performance.

Acknowledgments—We are grateful to Donald MacGregor for providing us with copies of his stimulus materials and to Gary McClelland and Charles Judd for advice on methodological and conceptual issues. We also thank Tom Stewart and an anonymous reviewer for their helpful comments on the manuscript, and Michele Nathan for her editorial assistance.

REFERENCES

Balzer, W.K., Doherty, M.E., & O'Connor, R., Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*, 410–433.

Brunswik, E. (1956). *Perception and the representative design of experiments.* Berkeley: University of California Press.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association, 68*, 361–368.

Cooksey, R. (1996). *Judgment analysis: Theory, methods and applications.* San Diego: Academic Press.

Hammond, K.R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62*, 255–262.

Hastie, R., Hammerle, O., Kerwin, J., Croner, C.M., & Herrmann, D.J. (1996). Human performance reading statistical maps. *Journal of Experimental Psychology: Applied, 2*, 3–16.

Hastie, R., & Rasinski, K.A. (1988). The concept of accuracy in social judgment. In D. Bar-Tal & A.W. Kruglanski (Eds.), *The social psychology of knowledge* (pp. 193–208). New York: Cambridge University Press.

Jarvenpaa, S.L., & Dickson, G.W. (1988). Graphics and managerial decision-making: Research based guidelines. *Communications of the ACM, 31*, 764–774.

MacGregor, D., & Slovic, P. (1986). Graphic representation of judgmental information. *Human-Computer Interaction, 2*, 179–200.

Pennington, N., & Hastie, R. (1993). A theory of explanation-based decision making. In G. Klein, J. Orasanu, R. Calderwood, & C.E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 188–204). Norwood, NJ: Ablex.

Singer, M., & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology, 30*, 1–38.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R.M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 5–27). Chicago: University of Chicago Press.

Stewart, T. (1988). Judgment analysis: Procedures. In B. Brehmer & C.R.B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 41–74). Amsterdam: Elsevier Science Publishers.

(RECEIVED 1/12/97; REVISION ACCEPTED 11/7/97)