Multiattribute Evaluation Reference Effects: A Reply to Barron and John

J. FRANK YATES

The University of Michigan

CAROLYN M. JAGACINSKI

Purdue University

In a recent paper, Yates and Jagacinski (1979) presented evidence for what was referred to as a "reference effect" in the process by which people evaluate multiattribute alternatives. Barron and John (1980) argue that it is inappropriate to characterize the observed phenomenon as a reference effect. We wish to take this opportunity to clarify what the issues are and what the data and analyses of Yates and Jagacinski and Barron and John really have to say about those issues.

ISSUE 1: THE PHENOMENON

The phenomenon in question is somewhat obscured in the Barron and John paper. Yates and Jagacinski speculated that multiattribute evaluation procedures requiring the evaluator to make reference to a baseline alternative that is uniformly worst on all relevant attribute dimensions are likely to be "problematic." It was argued that the problems rest in two things: (a) the difficulty of actually appreciating what such catastrophic alternatives really entail (because they are very rare) and (b) the likely difference in character of evaluations of alternatives similar to these catastrophic alternatives as compared to evaluations of alternatives that are not so "underwhelming." The second part of the argument was specialized to a claim that most preference structures one is likely to encounter are such that differences among alternatives close to the uniformly worst alternative matter very little to the evaluator, while equivalent differences among alternatives close to the uniformly best alternative matter a great deal. So, the phenomenon under consideration is the evaluation process over particular regions of an alternative or "consequence" space, in the language of Keeney and Raiffa (1976, p. 67). Yates and Jagacinski asserted that in regions near the uniformly worst alternative, evaluations should be ambiguous, but generally such that evaluation differences are minimal. In contrast, evaluations in other regions of the

The reported research was supported by Grant 27000 from the National Institute of Mental Health. Requests for reprints should be sent to: J. Frank Yates, 136 Perry Building, Department of Psychology, University of Michigan, Ann Arbor, MI 48104.

0030-5073/80/030375-09$02.00/0

space should be much more meaningful. In addition, evaluation differences among alternatives close to the alternative considered to be best on all relevant dimensions should be quite sizable.

ISSUE 2: EXPLANATIONS OF THE PHENOMENON

Why should the kinds of reference effects described above occur or not occur? That is, how might evaluations over particular areas of a consequence space be explained? This is the most important issue of all, from a theoretical standpoint.

Barron and John purport to provide an "explanation" of the effects in question in terms of multiplicative value functions of a given form. Such an approach might provide a description of the effects, but hardly an explanation. An analogy clarifies the point: Suppose an investigator collects a set of paired observations of two variables and then plots them. He observes that the plot is monotone increasing in a nice, regular pattern. He then fits a regression line to the points and finds that it characterizes the plot fairly well. Few of us would then be willing to say that the investigator has "explained" the relationship between the two variables; he has merely described it, albeit perhaps effectively and parsimoniously. Common experience and intuition tell us that the simplest type of well-behaved function that has the same form as the reference effects hypothesized by Yates and Jagacinski is a multiplicative function that "fans out." This is what Barron and John claim as their major point. This observation does not advance us much further than the hypothetical investigator who explains already known bivariate relationships by fitting regression lines.

We contend that a key to understanding reference effects is recognizing that any multiattribute evaluation situation involves not one, but three preference structures: (a) a "judged preference structure"; (b) a "normative preference structure"; and (c) an "experienced preference structure." An example highlights the distinctions. A particular car buyer is confronted with tradeoffs among the dimensions of style, operational economy, comfort, and price. If he is provided with an array of descriptions of cars in terms of these characteristics and asked to evaluate them, he does so using his judged preference structure. If he is asked to report how he feels he "should" evaluate such an array of car descriptions, his normative preference structure becomes operative. If he is allowed to actually see and "experience" each car and is then required to indicate his evaluations, his experienced preference structure is called into play.

There is no reason to expect all three of these preference structures to be identical. Indeed, certain systematic discrepancies ought to be anticipated, although the structures should rely on one another in particular ways. Experienced preferences are in some sense the ultimate criterion in decision making. Unfortunately, they are almost never discussed in the decision literature. One can therefore only speculate about the nature of experienced preference structures. For a particular attribute dimension i, $i = 1, \ldots, n$, suppose that the best conceivable level of the dimension is represented by B_i^*, while the worst conceivable level of the dimension is symbolized by W_i^*. The entire consequence space, what might be referred to as the "complete" consequence space, is then the product set $C^* = X_1^* \times X_2^* \times \ldots \times X_n^*$, where X_i^* is the set consisting of all conceivable levels of attribute dimension i. Consider the alternative $W^* = (W_1^*, W_2^*, \ldots, W_n^*)$, the uniformly worst alternative. Two things are likely to be true about W^*. First, W^* probably does not exist. While individually each of W_i^*, $i = 1, \ldots, n$, is likely to be achievable, their joint occurrence in many instances is not only highly improbable, but practically impossible. For instance, in the domain of cars, what manufacturer would offer a car that for even one individual is considered to be the ugliest, the most costly to run, the most wretched to ride in, and the most expensive to buy? The market would rapidly see to the extinction of such manufacturers. The second observation: Suppose that W^* is altered by changing one of W_i^*, $i = 1, \ldots, n$, to B_i^*, i.e., W^* becomes $W_{-i}^* = (W_1^*, \ldots, B_i^*, \ldots, W_n^*)$. How are the evaluations of W^* and W_{-i}^* likely to compare with each other? Most individuals one might query will find it hard to bring to mind a consequence space in which such evaluations will be substantially different from each other. (Try it!) Moreover, the likelihood of large differences in the evaluations of W^* and W_{-i}^* decreases as the dimensionality of the consequence space increases, i.e., as the number of relevant attribute dimensions increases. The dampening effect of the worst-level attributes grows as they become more numerous.

Consider the alternative $B^* = (B_1^*, B_2^*, \ldots, B_n^*)$. Alternatives such as B^* should not be nearly as rare as those similar to W^*, if for no other reasons besides market considerations. They are, nevertheless, likely to be scarce and inaccessible because they will be in great demand. Moreover, structural constraints will often make the existence of alternatives such as B^* actually impossible. For example, one cannot conceive of a car that is simultaneously the largest and the least expensive to operate. How might the evaluations of B^* and $B_{-i}^* = (B_1^*, \ldots, W_i^*, \ldots, B_n^*)$ compare? Again, we submit that while most individuals will find it easier to bring to mind consequence spaces where $B^* - B_{-i}^*$ evaluation differences are large than to think of spaces where $W^* - W_{-i}^*$ evaluation differences are large, they will still find it a fairly hard task. In other words, we suspect that experienced preference structures over complete consequence spaces are represented by evaluation surfaces that are fairly flat near alternatives such as W^* and B^*, but much steeper practically everywhere else. Moreover, there is little reason to expect such surfaces to generally have nice, regular forms perfectly describable by simple rules such as additive or multiplicative value functions.

In typical decision analyses, the relevant consequence space is not the complete one C^*, but rather a restriction of it, $C = X_1 \times X_2 \times \ldots \times X_n$, $i = 1, \ldots, n$. Restricted consequence spaces extend over attribute dimensions whose bounds are defined by levels W_i and B_i, $i = 1, \ldots, n$, respectively. W_i is not less preferred than W_i^*, and is generally preferred to W_i^*. It is that level of attribute dimension i that is least attractive among all plausibly-encountered levels in the given situation. B_i is defined in an analogous fashion. Generalizing from what is suggested above, one would expect that among the alternatives available in a practical decision situation, evaluation differences among alternatives nearest the uniformly worst alternative would be least, while differences among alternatives nearest the uniformly best alternative would be greatest. The latter alternatives, which are near $B = (B_1, B_2, \ldots, B_n)$ in C, are likely to be rather middling in C^*, i.e., in the region where the evaluation surface is steepest.

What can a decision maker rely on when he or she is asked to evaluate a description of a hypothetical alternative in a decision making experiment or a decision analysis? In other words, what is the basis for a person's judged preference structure? We contend that judged preferences and evaluations rest on three things. First, they are in part the evaluator's speculations about what his or her experienced preference structure is like. As such, judged preferences clearly can be "wrong" in the same way that all predictions or estimates can be wrong. Moreover, judged preferences and evaluations should be most in error for those types of alternatives with which the evalutor has had least experience. Thus, for instance, if a person who has only owned and driven compact cars all his life is asked to judge how he would evaluate an expensive, stylish, and well-built automobile, he might well be way off base. The second basis for judged evaluations is related to the first. We suspect that such evaluations depend heavily on the extent to which the descriptions of the alternatives are effective in getting the evaluator to psychologically "experience" each alternative. For instance, it is undoubtedly the case that pictures of particular automobiles will permit more accurate judgments of tradeoffs involving style than will verbal descriptions of those automobiles. The descriptions of alternatives in most decision analyses and experiments are remarkably impoverished. Accordingly, we cannot help wondering how seriously we should take the conclusions based on such exercises. Reading accounts of dialogues between decision analysts and their clients (e.g., in Keeney, 1977), one doubts that the client really appreciates what an alternative like $(W_1, W_2, \ldots, W_n)$ really means when the attribute dimensions are such things as fatalities, radioactive wastes, chronic health effects, and nuclear safeguards.

The final basis for judged evaluations and preferences is the person's normative preference structure. Many decision makers are likely to feel that there is a way they *should* make evaluations, regardless of how they actually *do* make evaluations—judged or experienced. For example, a car buyer might recognize that his tradeoffs between style and economy have a certain form: "When I see a beautiful car, I just say 'To hell with the economy.' " He might believe that those tradeoffs are not in his ultimate best interests, however, and seek to change them. Thus, when faced with a set of descriptions of alternatives, the evaluator is likely to moderate his or her judged evaluations to conform with the prescriptions of his or her normative preference structure. Discrepancies from such prescriptions would be considered errors of judgment.

In terms of reference effects, what this all means for the sorts of judged evaluations and preferences that form the cornerstone of decision analysis is the following: (1) Evaluation differences should be minimal near uniformly worst alternatives and largest near uniformly best alternatives, particularly when there are large numbers of attribute dimensions and when the descriptions of the alternatives are effective ones. (2) Evaluations of alternatives most similar to uniformly worst alternatives should contain a great deal of error, in the sense of being discrepant from what the evaluator's experienced preference structure might require. (3) Evaluations of all alternatives might well exhibit substantial error relative to the evaluation policy prescribed by the evaluator's normative preference structure. In particular, the sorts of regional evaluation surface characteristics implied in (1) might be contrary to the way the evaluator feels he or she really ought to judge things.

ISSUE 3: EVIDENCE CONCERNING THE PHENOMENON

The evidence cited in both the Yates and Jagacinski and the Barron and John papers is primarily concerned with Claim 1. Yates and Jagacinski reported the results of nomothetic analyses of 48 subjects' evaluations which were consistent with Claim 1. Barron and John drew conclusions on the basis of the responses of Fischer's (1976) 10 subjects. Barron and John reported that riskless responses (and those are the ones most comparable to the Yates and Jagacinski subjects' responses) of 7 of those subjects revealed no reference effects at all, while the responses of the remaining three subjects were said to suggest an effect opposite to that prescribed by Claim 1.

What should be made of the apparent discrepancy in results? First of all, when given the choice between conclusions based on the responses of 48 subjects and conclusions based on the responses of 10 subjects, most of us would be inclined toward the former. There were several important procedural differences between the Yates and Jagacinski and Fischer studies: (a) While Yates and Jagacinski's subjects made evaluations of alternatives described in terms of their status along 5 attribute dimensions, the alternatives in Fischer's study were characterized on only 3 dimensions. As indicated, the type of reference effect described in Claim 1 should increase in magnitude with the number of dimensions available to the subject. (b) All of Fischer's subjects evaluated alternatives defined by the same three attributes: salary, city, and type of work associated with prospective jobs. In contrast, the university courses evaluated by Yates and Jagacinski's subjects could be defined by any 5 out of a collection of 52 different attribute dimensions. On these grounds alone one would lean toward accepting the generality of the results of the Yates and Jagacinski studies. (c) Each subject in the Yates and Jagacinski studies was required to actually bring to mind and write out descriptions of concrete exemplars of the relevant levels of each of the 5 attribute dimensions characterizing the alternatives he or she evaluated. On the other hand, such a procedure was required of Fischer's subjects for only 1 dimension, type of work, Thus, it is quite possible that the subjects in the Fischer study did not have as full an appreciation of what was actually entailed by each of the alternatives they evaluated as did the Yates and Jagacinski subjects. If a subject does not really think about what various alternatives mean, he or she should be relatively more inclined to evaluate them in an additive fashion. It is easier to perform evaluations that way; one does not have to pay much attention to what he or she is doing. (d) The best and worst levels of two of Fischer's 3 attribute dimensions, salary and city, were set arbitrarily by the investigator. In contrast, Yates and Jagacinski's subjects were instructed to define the best and worst levels of the attribute dimensions describing the courses they evaluated as those one could expect among elective courses at a rather large state institution, the University of Michigan. Thus, one would expect the best and worst levels of the attribute dimensions in the Yates and Jagacinski studies to be much farther apart in subjective value than those in the Fischer study, thus enhancing the chances of the types of reference effects of Claim 1.

There is one final aspect of the Barron and John analyses that may be worrisome. Barron and John's analysis of Fischer's data relied upon a holistic assessment procedure developed by Barron and Person (1979). A key feature of that technique is the estimation of the *K* parameter in Keeney's (1974) or Dyer and Sarin's (1979) multiplicative utility and value functions. That parameter is estimated on the basis of two judgments defined by only one of the attribute dimensions available to the subject. Barron and John classified subjects' utility and value functions on the basis of such *K* estimates. Unfortunately, they did not report just what those estimates were. One of the difficulties of this procedure is that one can easily obtain different estimates for the same subject, depending on the dimension used in the procedure. When that occurs, what does one conclude? When does one attribute discrepancies to random error and when does one conclude that the discrepancies signal that neither the additive nor the multiplicative model describes the subject's judgments? We applied the K estimation technique of Barron and Person to our subjects' responses and found substantial differences among estimates of K based on different dimensions. This result forces us to question Barron and John's conjecture that those subjects' preference structures conform to a multiplicative model with $K > 0$, even though among simple algebraic models that model provides the closest fit.

The data cited by Barron and John as well as those of Yates and Jagacinski have essentially no bearing on Claim 2, the assertion concerning the discrepancy between judged and experienced evaluations. As indicated above, the decision literature (with the exception of the social psychological decision literature, e.g., Janis & Mann, 1977) is devoid of attempts to even examine such discrepancies. A test of the assertion requires a longitudinal study in which judged and experienced evaluations of particular alternatives over a large consequence space are compared to each other. To the best of our knowledge, this sort of study has never been done. The closest to it are previously unreported studies conducted by ourselves. In one study, subjects evaluated, then enrolled in and completed, and finally evaluated again university elective courses. In the other study, subjects similarly provided preliminary and retrospective evaluations of films. In both studies, initial and final evaluations bore little relation to each other. There are several reasons these studies do not provide a good test of Claim 2. Nevertheless, they do not make one optimistic that judged preference structures generally conform well to experienced preference structures.

The results of Yates and Jagacinski constitute evidence directly relevant to the third claim outlined above, the hypothesis that discrepancies between judged and normative evaluations might be rather common. Indeed, the primary data of the Yates and Jagacinski studies amount to graphic indications of such discrepancies. Subjects were instructed to select attribute dimensions such that their evaluations of alternatives varying along those dimensions would be unaffected by the status of the alternatives with respect to the remaining dimensions. The subjects' evaluations clearly did not conform to that requirement.

CONCLUSIONS AND IMPLICATIONS

The jury is still out. It may well be the case that the propositions we put forth are wrong. They are testable, however, and in due time *will* be subjected to tests beyond those reported by Yates and Jagacinski and Barron and John. For the moment, the weight of the evidence seems to be consistent with those propositions.

Suppose the hypothesized effects do hold up upon rigorous testing. Consider their implications for decision analysis procedures. Most simple rating methods such as Edwards' SMART technique (Gardiner & Edwards, 1975) ignore issues of attribute interactions and assume that preference structures are additive. The reference effects hypothesized here suggest that such an assumption is surely wrong, at least in regard to judged and experienced preference structures. It is an open question whether a given decision maker's normative preference structure is additive, i.e., whether he or she would like for his or her preferences to be describable by an additive model. Thus, it is probably wise for users of methods such as SMART to incorporate procedures intended to find out what the decision maker's normative preference structure is like and to test the sensitivity of resultant decisions to the form of the normative preference structure.

Theoretically precise techniques such as those advocated by Keeney and Raiffa (1976) include routines for checking the consistency of the decision maker's judged preference structure with various conditions required for representing such structures by simple algebraic models. So, presumably, the types of effects hypothesized here would manifest themselves at the condition-testing stage of a decision analysis. Thus, we should expect that more often than not, conditions for simple algebraic models will not be met. Even when they *are* apparently met, however, one should be careful to see whether the decision maker really appreciates what alternatives with large numbers of worst-level attributes really mean or whether he or she is simply taking the line of least resistance and only verbally acknowledging agreement with such principles as preferential independence. Our view of the decision problem suggests that the very notion of modeling a decision maker's judged or even experienced preference structure might miss the major point of attempting to aid decision makers. The typical decision maker is not interested in modeling how he or she *does* make decisions, but rather how he or she *should* make decisions. Thus, decision analysts should reorient their basic emphasis away from the modeling of judged preference structures to the modeling of normative preference structures.

REFERENCES

Barron, F. H., & John, R. Reference effects: A sheep in wolf's clothing. *Organizational Behavior and Human Performance*, 1980, 25, 365–374.

Barron, F. H., & Person, H. B. Assessment of multiplicative utility functions via holistic judgments. *Organizational Behavior and Human Performance*, 1979, 24, 147–166.

Dyer, J. S., & Sarin, R. K. Measurable multiattribute value functions. *Operations Research,* 1979, 27, 810–822.

Fischer, G. W. Multidimensional utility models for risky and riskless choice. *Organizational Behavior and Human Performance,* 1976, 17, 127–146.

Gardiner, P. C., & Edwards, W. Public values: Multiattribute-utility measurement for social decision making. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes.* New York: Academic Press, 1975.

Janis, I. L., & Mann, L. *Decision making.* New York: The Free Press, 1977.

Keeney, R. L. Multiplicative utility functions. *Operations Research,* 1974, 22, 22–34.

Keeney, R. L. The art of assessing multiattribute utility functions. *Organizational Behavior and Human Performance,* 1977, 19, 267–310.

Keeney, R. L., & Raiffa, H. *Decisions with multiple objectives: Preferences and value tradeoffs.* New York: Wiley, 1976.

Yates, J. F. & Jagacinski, C. M. Reference effects in multiattribute evaluations. *Organizational Behavior and Human Performance,* 1979, 24, 400–410.

RECEIVED: December 20, 1979