ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 24, 400–410 (1979)

# Reference Effects in Multiattribute Evaluations

J. FRANK YATES AND CAROLYN M. JAGACINSKI

*The University of Michigan*

The studies reported here examined the differences in evaluative ratings given to pairs of multiattribute alternatives that were, respectively, at the "worst" and "best" plausible levels on one target dimension, but at the same constant level on the remaining dimensions. Results of the first study indicated that evaluative differences were significantly greater when nontarget dimensions were held constant at their "typical" (usually intermediate) levels rather than at their worst levels. In the second study such differences were larger still when the nontarget dimensions were maintained at their best levels. The results are interpreted as evidence of a systematic violation of requirements for linear representations of subjective multiattribute evaluation policies. Supplementary data analyses and simulations of decision situations were conducted to assess the implications of the revealed reference effects for decision analysis procedures. Those results suggested that ignoring reference effects and assuming a linear model lead to prescriptions that may well be inappropriate.

Almost all practical evaluation and decision situations require one to make trade-offs among two or more salient dimensions that characterize the alternatives, e.g., an employee's initiative vs his/her reliability, or an automobile's appearance vs its performance. Such problems involving multiple objectives or considerations have attracted a great deal of attention in the evaluation and decision literatures of late (Keeney & Raiffa, 1976; Slovic, Fischhoff, & Lichtenstein, 1977; Zedeck & Kafry, 1977). Most of the models for describing and prescribing appropriate behavior in such multiattribute evaluation or decision situations have been linear or, occasionally, multiplicative. To date, systematic research and practice in the field suggest that in a wide variety of circumstances, subjects either make their judgments in a fashion consistent with a linear model or express agreement with principles compatible with such a model.

An increasingly popular class of techniques for representing people's evaluation policies requires the subject to make judgments of how much changing an alternative from its "worst" to its "best" plausible levels on its respective attribute dimensions would affect the overall evaluation of the alternative (Edwards, Guttentag, & Snapper, 1975; Keeney & Raiffa,

This research was supported by Grant 27000 from the National Institute of Mental Health. The results of the research were presented at the convention of the American Institute for Decision Sciences, St. Louis, November 1978. Requests for reprints should be sent to J. Frank Yates, 136 Perry Building, Department of Psychology, University of Michigan, Ann Arbor, MI 48104.



0030-5073/79/060400-11$02.00/0

1976). The archetypal worst-to-best procedure provides the subject with a reference alternative that is at the worst plausible level on all relevant dimensions. Each of the remaining alternatives is identical to the reference alternative except for one target dimension on which it is at the best plausible level. The differences in the evaluations given to the reference alternative and the remaining alternatives are assumed to reflect the relative importance of the respective attribute dimensions in the subject's evaluation policy.

Symbolically, the procedure can be represented as follows: Let a given alternative $A_j$ be expressed in terms of its significant attributes $1-k$ by the profile or vector $(X_{j1}, X_{j2}, \ldots, X_{jk})$, where $X_{ji}$ is an index of the status of $A_j$ on attribute dimension $i$. Let $M_i$ and $m_i$ represent, respectively, appropriately scaled indexes of the levels of attribute dimension $i$ that are best and worst. A level is "best" if the subject can envision it occurring and it is the most highly preferred of all such conceivably achievable levels. The operational meaning of "worst" is analogous. The following set of $k + 1$ profiles would provide the evaluation policy information desired:

$A_0 = (m_1, m_2, \ldots, m_k)$ (worst level on all dimensions; the reference alternative)
$A_1 = (M_1, m_2, \ldots, m_k)$ (best on dimension 1, worst on rest)
$A_2 = (m_1, M_2, \ldots, m_k)$ (best on dimension 2, worst on rest)
. . .
. . .
. . .
$A_k = (m_1, m_2, \ldots, M_k)$ (best on dimension $k$, worst on rest)

Suppose that the subject's evaluation policy is linear, i.e., the subject's scaled evaluation of alternative $A_j$ is given by $S_j = \sum_{i=1}^{k} w_i X_{ji}$. Then we have the following set of critical differences:

$$D_1 = S_1 - S_0 = w_1 (M_1 - m_1)$$
$$D_2 = S_2 - S_0 = w_2 (M_2 - m_2)$$
. . .
. . .
. . .
$$D_k = S_k - S_0 = w_k (M_k - m_k)$$

Or, in general, $D_j = S_j - S_0, j = 1 - k$. Provided that the best and worst levels are properly scaled, $M_i$ and $m_i$ can be assumed to be constants with respect to $i$, with $M_i > m_i$. Thus, the differences $D_j, j = 1 - k$, provide scale values of the various attribute weights, the $w_i$'s.

Now, if the linear model is truly representative of the subject's evaluations, the reference alternative one uses for deriving evaluation differences and, thereby, inferring dimensional weights should be of little con-

sequence. However, informal observation of subjects making the kinds of judgments demanded by the standard worst-to-best procedure suggested that use of the uniformly worst reference might indeed be quite problematic. The required judgments are not only difficult for subjects to make, they are disconcerting as well. An alternative that is worst on all attribute dimensions is psychologically devastating. It is, first of all, very hard to envision and internalize the meaning of an alternative that is so consistently and catastrophically bad. Moreover, intuition suggests that a subject is not likely to think that adjusting only one of the attribute dimensions of such an overwhelmingly unattractive option can improve its worth substantially.

An alternative judgment routine for inferring weights that should be less susceptible to the problems suggested above requires the subject to conceptualize a "typical" level of an attribute dimension, $t_i$, in addition to the best and worst levels. The "typical" level can be defined as that level the subject thinks an attribute dimension is most likely to assume, i.e., a modal level. Then, an appropriate set of profiles for deriving dimensional weights would include the following 2 $k$ alternatives.

$A_{1M} = (M_1, t_2, \ldots, t_k)$ (best on dimension 1, typical on rest)
$A_{1m} = (m_1, t_2, \ldots, t_k)$ (worst on dimension 1, typical on rest)
$A_{2M} = (t_1, M_2, \ldots, t_k)$ (best on dimension 2, typical on rest)
$A_{2m} = (t_1, m_2, \ldots, t_k)$ (worst on dimension 2, typical on rest)
. .
. .
. .
$A_{kM} = (t_1, t_2, \ldots, M_k)$ (best on dimension $k$, typical on rest)
$A_{km} = (t_1, t_2, \ldots, m_k)$ (worst on dimension $k$, typical on rest)

Evaluations of these profiles should yield, according to the linear model, the same $D_j$ as before, via a different route:

$$D_1 = S_{1M} - S_{1m} = w_1 (M_1 - m_1)$$
$$D_2 = S_{2M} - S_{2m} = w_2 (M_2 - m_2)$$
. . .
. . .
. . .
$$D_k = S_{kM} - S_{km} = w_k (M_k - m_k)$$

Or, in general, $D_j = S_{jM} - S_{jm}$.

## STUDY 1

The specific issues addressed by the first study reported here were these: First, consistent with previous informal observations and hypotheses about reference effects, are subjects' actual dimensional evaluation

differences generally greater when they are made relative to a typical rather than a worst reference alternative? Second, assuming that the answer to the first question is affirmative, would relative importance weights derived by the alternative elicitation procedures be comparable?

*Method*

*Subjects.* All 24 of the subjects used in the study were students at the University of Michigan who had been enrolled for at least one term. The subjects were paid a flat rate for participating in the study.

*Stimuli.* Stimuli consisted of profiles of hypothetical, nonrequired, university courses which varied along five attribute dimensions. Each subject specified the five dimensions which defined all the courses that he/she was to judge. The selection of attributes was facilitated by requiring the subject to survey descriptions of an exhaustive list of 52 course dimensions previous studies indicated affect students' evaluations of university courses. The only restrictions placed on attribute dimensions employed by a subject were that they be intuitively independent of one another, in both the probabilistic and value senses of the term. After the subject specified his/her set of five independent course attribute dimensions or "factors," as they were termed, he/she was told to "create your own scale ranging from the worst level to the best level" on each dimension. "Best," "worst," and "typical" were defined as described previously. To make certain that each subject had a concrete interpretation of each dimension, he/she was required to write down real or imaginary instances of courses or instructors exemplifying his/her notions of the best, worst, and typical levels of each factor. Thus, for example, for the dimension, "Instructor Friendliness," one subject wrote the following as exemplars for the best, typical, and worst levels, respectively: "When he's eager to teach and help others learn;" "Paul Newhouse;"[1] and "When his friendliness covers up his stupidity." The best, typical, and worst levels of "Marketable Skills" for another student were described, respectively, as: "Gives some kind of knowledge useful in a job;" "No knowledge useful in a job (History);" and "Class has nothing to do with the student's future." Each subject's exemplars of best, worst, and typical levels were arranged in columnar fashion by dimension on a master sheet. As implied by the technique outlined above, each profile evaluated by subjects was described in terms of its status on each dimension as best, typical, or worst. So, corresponding to each of the required profiles was a specially constructed response sheet that fit over the master sheet. The response sheet had windows cut into it so that it exposed the best, typical, and worst level exemplars representative of the desired profile. On a continuous Likert-type scale with anchors "Maximum Satisfaction," "Indiffer-

[1] Not the actual name cited by the subject.

ence," and "Maximum Dissatisfaction," the subject could indicate his/her anticipated evaluation of a course represented by the given profile.

*Procedure.* The purpose of the larger study of which the present study was a part was explained to the subject. The scaling procedures were then described and the subject completed his/her master sheet of best, worst, and typical exemplars. The profiles the subject was required to evaluate included all 10 demanded by the typical baseline procedure. Three of those needed in a complete version of the worst baseline procedure were included also, the profile that was worst on all dimensions and those that were best on the second and fourth dimensions, going from left to right on the subject's response sheet. These 13 profiles were presented to the subject in random order for evaluation.

*Results and Discussion*

Subjects' evaluations were coded on a 20-point scale, with high scores corresponding to favorable assessments. Figure 1 illustrates the results of primary interest. As anticipated, the evaluation differences along the two dimensions that permitted comparisons were substantially smaller when judgments were made relative to a worst rather than a typical reference alternative ($F(1,23) = 19.33$, $p < .01$). There was no statistically significant effect of attribute dimension display position or interaction of position and reference. So, there seems to be a distinct bias in judgment that is induced by the reference provided to the subject. Clearly, the pattern of

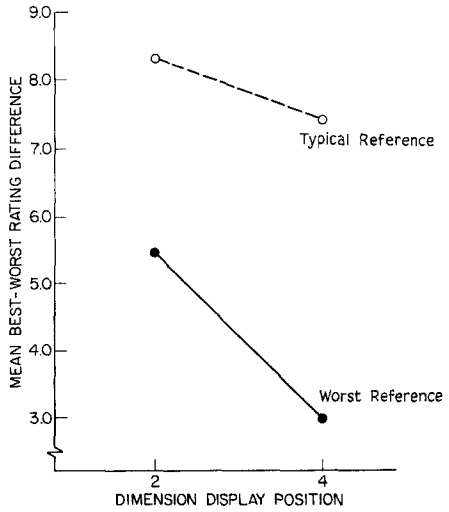


FIG. 1. Mean best–worst rating differences by reference alternative and dimension display position, Study 1.

responses provided by the subjects is incompatible with a linear evaluation model.

Suppose, as is commonly done, one wanted to ignore the shortcomings of the linear model and operate "as if" it does a decent job of describing evaluation policies in a given practical setting. How should one derive dimensional weights? Given that a profile that is typical on most attribute dimensions is closer to most subjects' real-world experience than one that is worst on most dimensions, it seems that deriving weights relative to a typical reference alternative would be more desirable. But, would the reference really matter? In practice, dimensional weights are ordinarily normalized to sum to 1 or 100. Given such normalization procedures, there is no formal necessity that the bias identified here should lead to different normalized derived weights via typical and worst reference procedures. It is conceivable that the type of reference alternative employed simply expands or contracts dimensional evaluation differences by a constant factor. If so, normalized weights would be identical, regardless of the reference.

The ratios of critical differences defined by the respective references permit some insight into the issue of normalized weight discrepancies. The required ratios for each subject were $R_2 = D_{2t}/D_{2w}$ and $R_4 = D_{4t}/D_{4w}$, where $D_{jt}$ is the evaluation difference for attribute dimension $j$, given a typical reference, and $D_{jw}$ is the corresponding difference, given a worst reference. If the reference bias amounts to a constant factor expansion or contraction, $R_2$ and $R_4$ should be the same. The mean ratios were, respectively, 2.50 and 4.86, a discrepancy that is marginally significant statistically ($t(17) = -1.84$, $p < .09$). (Note: There were fewer cases in this analysis because instances of zero $D_{jw}$'s were excluded.) So, it is not clear whether in practice the reference alternative will affect derived normalized weights. There is some reason to suspect, however, that it might.

## STUDY 2

The second study reported here was intended to pursue several issues raised by Study 1. The first question was simply whether the major result of that study could be replicated. The second issue was whether the reference bias would exhibit itself across more than just two of the dimensions displayed to the evaluator. The final, and perhaps most interesting, question was whether the reference bias extends to evaluation differences relative to a best reference alternative, i.e., one that is at the best plausible level with respect to all relevant attribute dimensions.

### *Method*

The method for Study 2 was essentially the same as that for Study 1, with the exceptions noted. Again, 24 subjects provided complete and usable evaluation responses. Altogether, each subject rated the satisfac-

toriness of 22 profiles of hypothetical courses. There were, first of all, the 2 extreme reference alternatives that were, respectively, at the worst and best levels on all attribute dimensions. Five profiles were worst on all attribute dimensions except one. Each of those alternatives was, respectively, at the best level on the remaining dimension. There was a complementary set of 5 alternatives that were each best on four dimensions and worst on the remaining dimension. Finally, there was a set of 10 alternatives that were each at the typical level on four dimensions and either best or worst on the last dimension.

*Results and Discussion*

Figure 2 displays the mean rating differences of concern. Again, there is a distinct reference effect ($F(2,46) = 14.94, p < .01$). Overall mean differences relative to the best and typical reference alternatives were both larger than that relative to the worst reference alternative ($p < .05$ for both comparisons, Newman–Keuls tests). Similarly, the overall mean difference relative to the best reference alternative was larger than that relative to the typical reference alternative ($p < .05$, Newman–Keuls test). There was no statistically reliable effect of dimension display position nor a significant interaction of display position and reference.

Thus, it appears that the effect of reference alternatives on multiattribute evaluations is indeed a replicable phenomenon. For the most part, it emerges with the same degree of strength across all attribute dimensions

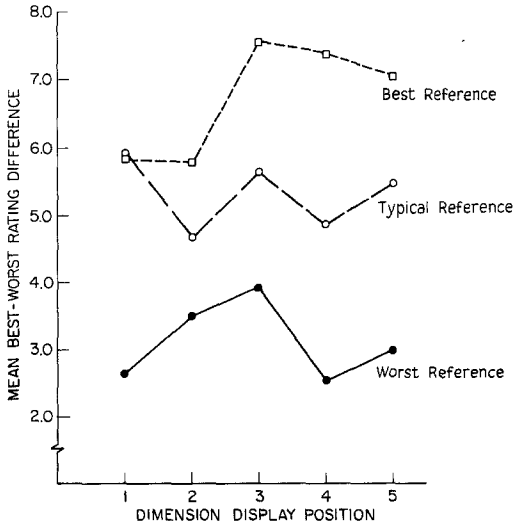


FIG. 2. Mean best–worst rating differences by reference alternative and dimension display position, Study 2.

available to the evaluator. The character of the effect is such that the evaluative significance of a particular dimension is greater the more attractive the given alternative happens to be on all the remaining dimensions.

## GENERAL DISCUSSION

What do the results of the present studies mean? From a purely psychological point of view, they identify a rather interesting feature of the way people respond to distinctions among objects. They suggest that we are not terribly responsive to differences among evaluative targets that are generally unattractive. In contrast, those same differences among targets that are generally pleasing induce much greater effects on our overall impressions. It is not obvious why this phenomenon occurs, although a couple of plausible explanations should be pursued in further studies. One possibility is that the observed reference effects reflect our sensitivity to changes in status. We may be much more strongly affected by deterioration in our status than by enhancement of our current state of affairs. (Deviations from a best reference necessarily constitute reductions in status, while deviations from a worst reference are necessarily increases in status.) Conceptually, this notion is something of a parallel to risk aversion, but under conditions of certainty (cf. Raiffa, 1968; or Winkler, 1972). One approach to testing this hunch would involve comparing differences between evaluations of profile pairs of the form $(t_1, \ldots, t_i, \ldots, t_k)$ and $(t_1, \ldots, M_i, \ldots, t_k)$ to differences in judgments of pairs like $(t_1, \ldots, t_i, \ldots, t_k)$ and $(t_1, \ldots, m_i, \ldots, t_k)$, where, as before, $M_i$, $m_i$, and $t_i$ correspond, respectively, to best, worst, and typical attribute levels. $(t_1, \ldots, t_i, \ldots, t_k)$ vs $(t_1, \ldots, M_i, \ldots, t_k)$ evaluative differences would index the effects of status improvements, while $(t_1, \ldots, t_i, \ldots, t_k)$ vs $(t_1, \ldots, m_i, \ldots, t_k)$ evaluative differences would reflect responses to status reductions where changes are not necessarily constrained in one direction or the other.

A second potential explanation for the revealed reference effects is perhaps more interesting than the first in terms of fundamental decision processes. The effect may be a consequence of the step-by-step operations decision makers execute when they choose among multiattribute alternatives. It is not unreasonable to propose that the choice procedure occurs something like the following: The decision maker scans the given pool of alternatives several times. Initial scans are intended to eliminate alternatives that are "obviously" so bad that the decision maker cannot envision them ultimately being chosen as the most satisfactory of the available options. Almost certainly included among such alternatives that are eliminated early are those that are at the worst level on several of the relevant attribute dimensions. Successive scans of the alternative pool

would eventually lead to the discarding of all alternatives except those that are not dominated, i.e., are not less attractive than any other alternative on all relevant dimensions. It is only within this final "efficient set" of alternatives that the decision maker must give serious consideration to the relative effects of differences along respective attribute dimensions. The decision maker's efficient set should contain a high concentration of alternatives that are best on one or more attribute dimensions. While distinctions among generally unappealing alternatives are not deliberated carefully, if they are noted at all, such distinctions will be taken seriously indeed among highly attractive options; they are the only basis on which a "rational" choice can be made. Assuming that subjects generalize their judgment dispositions from choice to evaluation situations, the reference effects observed in the present studies would follow directly; evaluation differences should be greatest among generally attractive alternatives. A number of techniques for studying the details of people's decision processes have been suggested recently (see, for example, Payne, Braunstein, & Carroll, 1978; Svenson, 1979). The amount of time subjects take to evaluate various types of alternatives would provide a means of testing the proposed process explanation of reference effects. If the hypothesis is correct, judgments of attractive profiles like $(M_1, \ldots, t_i, \ldots, M_k)$ should be made much more slowly than those of options like $(m_1, \ldots, t_i, \ldots, m_k)$, where the notation is as defined previously.

What about the practical implications of the results? The results clearly suggest that in many circumstances the linear model does not provide a good representation of people's value functions under certainty. Although the present studies did not directly investigate judgments under uncertainty, there are direct analogs of the reference effects demonstrated here that would likely emerge in situations in which the alternatives are risky. In particular, it should not be surprising to discover pervasive and systematic violations of both preferential and utility independence, the two critical conditions for additive and multiplicative utility functions (cf. Keeney, 1974, 1977).

It is unclear at this time just how seriously the consequences of such judgment peculiarities would be for decision analyses. If an analyst must model the decision maker's preference structure by anything other than an additive or multiplicative representation, the complexity of the analysis can very rapidly become quite unwieldy and its usefulness questionable. On the other hand, if the analyst chooses to ignore reference effects and proceed as if the additive or multiplicative representations are appropriate, the resulting distortion of prescribed actions might be substantial enough to lead to costly decision errors.

To gain some impression of what the practical consequences of such an approach *might* be, three sets of normalized dimensional weights appro-

priate for a linear representation were derived for each subject in Study 2. The weights were based, respectively, on rating differences relative to the worst, typical, and best reference alternatives. Fifty samples of five randomly generated orthogonal profiles were assembled. The weights were applied to the profiles in each collection via a linear equation. The "choice" prescribed by each set of weights was taken to be the profile with the highest score as computed with the linear model incorporating the respective weights. The mean prescriptive agreement percentage of the worst and typical weights was 74%, of the worst and best weights 72%, and of the typical and best weights 76%. Over all comparisons, the lowest prescriptive agreement percentage was 34%, while the highest was 94%. It would be desirable to know the extent to which the observed inconsistency of prescriptions is due to a lack of reliability in subjects' judgment reports. Future studies should incorporate procedures for assessing degrees of reliability. It is unlikely, however, that unreliability can account for the inconsistency completely. The revealed reference effects seem much too strong and systematic for that possibility.

At first blush, prescriptive agreement percentages of 70% or more might look pretty good. Just how bad 25%–30% disagreement is, of course, depends on the cost of an error in a given decision situation. That cost may be negligible or it may be monumental. It might also be noted that 70% prescriptive agreement is probably unrealistically high to anticipate for most practical circumstances. In the real world, attribute dimensions are likely to be negatively correlated rather than independent as in our simulation. As suggested by McClelland (Note 1), this is to be expected because nature generally trades off good things for bad and because the serious contenders in choice situations typically involve even stronger tradeoffs. So, for instance, it is unlikely one will find a car that is both large and gets outstanding gas mileage. McClelland (Note 1) and Newman, Seaver, and Edwards (Note 2) have demonstrated that under such conditions of negatively correlated attribute dimensions, the prescriptive consistency of different linear weighting schemes would be very slight. McClelland also reports a simulation suggesting that when two sets of linear weights do not have the same rank ordering, the difference in value or utility of the prescriptions of the two weighting schemes is particularly large. Study 2 subjects' weights derived relative to worst, typical, and best reference alternatives seemed to differ in a number of ways. For example, their variabilities were significantly different ($F(2,46) = 4.38$, $p < .05$), with weights based on worst reference alternatives being least uniform across attribute dimensions. Most important, however, while each subject's sets of weights did not differ radically from one another, it was extremely unusual for all three sets of weights to have the same rank ordering by magnitude. This occurred for only 2 of the 24

subjects. Thus, it seems that the practical consequences of reference effects may well be worth serious consideration in a given applied setting.

## REFERENCES

Edwards, W., Guttentag, M., & Snapper, K. A decision-theoretic approach to evaluation research. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research, Vol. 1*. Beverly Hills, Calif.: Sage, 1975.

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.

Payne, J. W., Braunstein, M. L., & Carroll, J. S. Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior and Human Performance*, 1978, **22**, 17–44.

Raiffa, H. *Decision analysis*. Reading, Mass.: Addison–Wesley, 1968.

Slovic, P., Fischhoff, B., & Lichtenstein, S. Behavioral decision theory. *Annual Review of Psychology*, 1977, **28**, 1–39.

Svenson, O. Process descriptions of decision making. *Organizational Behavior and Human Performance*, 1979, **23**, 86–112.

Winkler, R. L. *An introduction to Bayesian inference and decision*. New York: Holt, Rinehart, & Winston, 1972.

Zedeck, S., & Kafry, D. Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance*, 1977, **18**, 269–294.

## REFERENCE NOTES

1. McClelland, G. H. *Equal versus differential weighting for multiattribute decisions: There are no free lunches*. University of Colorado, Center for Research on Judgment and Policy, Report No. 207, 1978.
2. Newman, J. R., Seaver, D. A., & Edwards, W. *Unit versus differential weighting schemes for decision making: A method of study and some preliminary results*. University of Southern California, Social Science Research Institute, Technical Report No. SSRI 76-5, 1976.

RECEIVED: November 22, 1978