Expressing Preferences in a Principal-Agent Task: A Comparison of Choice, Rating, and Matching

Joel Huber

Fuqua School of Business, Duke University

Dan Ariely

Sloan School of Management & the Media Laboratory, MIT

and

Gregory Fischer

Fuqua School of Business, Duke University

One of the more disturbing yet important findings in the social sciences is the observation that alternative tasks result in different expressed preferences among choice alternatives. We examine this problem not from the perspective of an individual making personal decisions, but from the perspective of an agent trying to follow the known values of a principal. In two studies, we train people to evaluate outcomes described by specific attributes and then examine their ability to express these known values with three common tasks: ratings of individual alternatives, choices among triples of alternatives, and matching pairs of alternatives to indifference. We find that each preference assessment method has distinct strengths and weaknesses. Ratings are quick, robust at following known values, and are perceived as an easy task by respondents. However, because ratings require projection to an imprecise response scale, respondents have difficulty when applying them to more complex preference structures. Further, they place too much weight on negative information, a result that is consistent with reference-dependent loss aversion. Choice is perceived as the most realistic task and the one about which people feel the most confident. However, choices exhibit the most

Address correspondence and reprint requests to Joel Huber, Fuqua School of Business, Duke University, Durham, NC 27708-0120. E-mail: Joel.Huber@duke.edu.

66

negativity, which, in addition to flowing from the same perceptual bias of ratings, may be exacerbated by a screening strategy that excludes alternatives possessing the lowest level of an attribute. Finally, the matching task takes the most time and is perceived to be the most difficult. It shows minimal biases, except for one glaring flaw, a substantial overweighting of the matching variable. This bias is consistent with a well-known compatibility bias and suggests that agents can learn to use a matching task appropriately for all attributes except the matching variable itself. The article concludes with a discussion of the theoretical mechanisms by which these biases infiltrate different elicitation modes and a summary of managerial implications of these results.   © 2001 Elsevier Science

Research in the field of judgment and decision making has generated convincing evidence that people construct their preferences in the light of demands produced by the situation and the response task (Payne, 1982; Kahneman & Tversky, 1984; Slovic, 1995). As predicted by this constructive view of preferences, different elicitation tasks evoke systematically different preferences. For example, studies of "task effects" have clearly demonstrated that a matching task, specifying the amount of an attribute required to make alternatives equal, results in quite different preference orderings than choice (Tversky, Sattath, & Slovic; 1988, Fischer & Hawkins, 1993; Hawkins, 1994; Ordóñez, Mellers, Chang, & Roberts, 1995). Related research shows that rating is different from choice (Bazerman, Loewenstein, & White, 1992; Fischer & Hawkins, 1993; Schkade & Johnson, 1989; Nowlis & Simonson, 1997; Delquié, 1993; Ahlbrecht & Weber, 1997) and that matching is different from rating (Fischer & Hawkins, 1993; Hsee, 1996). While there has been active debate on the mechanisms behind these phenomena, there is little doubt that the preferences revealed depend on the questions asked. In this work, we examine whether similar preference shifts occur for agents in the context of three tasks—ratings of individual options, matching of pairs to indifference, and choice among triples. Studying these methods within an agent task is important because it can tell us not only how the methods differ from each other but also how they differ from the true preference structure which the agent seeks to emulate.

Our use of an agent is similar to its use in multiple cue probability learning (Hammond, Summers, & Deane, 1973), but with a different goal. Whereas that stream of research is concerned with how people *learn* probabilistic cues in the environment, our focus is on the consistency and biases associated with human ability to *transmit* known values using different preference elicitation tasks. A strong point of difference is that the multiple probability cue learning paradigm requires subjects to infer policy from noisy feedback on the evaluation of profiles. In our agent tasks, there is no need to learn the "partworth values"— they are always displayed with graphs such as shown in Fig. 2. At issue is the extent to which people can correctly apply a given set of partworths under different tasks. Our approach shares kinship with the work of Klein and Bither

(1987), who used an analogous agency task to explore cutoff use in simplifying choices. Similarly, Stone and Kadous (1997) used an agent task to estimate the impact of ambient affect and task difficulty on choice accuracy. In contrast to both of these articles, our focus is less on estimating accuracy than on identifying consistent biases that arise.

An additional advantage of the agent task is that it isolates those biases that occur in the expression of preferences. If we consider three stages in the general value judgment problem as comprehending the information, understanding appropriate trade-offs, and expressing those trade-offs through a specific task, then our study focuses on the last stage. Our task thereby provides an upper bound on decision-makers' ability to express value through different tasks.

THEORETICAL DIFFERENCES AMONG THE TASKS

What kinds of biases would one expect to emerge among the choice, rating, and matching tasks studied here? The theoretical framework that we adopt is that people develop strategies that enable them to minimize effort while preserving accuracy (Bettman, Johnson, & Payne, 1990). As Peter Wright (1974) and Hillel Einhorn (1971) suggested more than 25 years ago, simplification can be achieved by focusing on the more important pieces of information. It is useful to distinguish between two ways in which this simplification can occur. First, *attribute focusing* occurs when more important attributes receive exaggerated attention. By contrast, *level focusing* occurs when it is the levels within attributes that get exaggerated attention.

Attribute Focusing

Attribute focusing minimizes effort by ignoring less important attributes. Russo and Dosher (1983) called this process *dimensional reduction*. To illustrate the way attribute focusing would be realized, imagine a target pattern of partworth values for ski trips such as those shown in Fig. 1A. In this hypothetical case, the full range of *price* is most important with 45% of the sum of the ranges of the other attributes, followed by *ski slope quality* with 35%, and then *probability of good snow* with 20%. Attribute focusing, pictured in Fig. 1B, increases the weight of the most important attribute, *price*, by 22%, while decreasing the weight of *slope quality* and *snow probability* by 14 and 25% respectively. Two mechanisms, prominence and scale compatibility, have been identified as leading to attribute focus.

The prominence effect reflects the empirical generalization that people are more likely to prefer an alternative that is superior on the more prominent attribute when making choices than when making judgments (Tversky, Sattath, & Slovic, 1988; Fischer & Hawkins, 1993; Hawkins, 1994). Contrasting Fig. 1A and 1B, the prominence effect predicts a greater slope to the most important attribute, *price*, relative to the other two. Scale compatibility is a second well-known attribute-focusing process in which people give greater
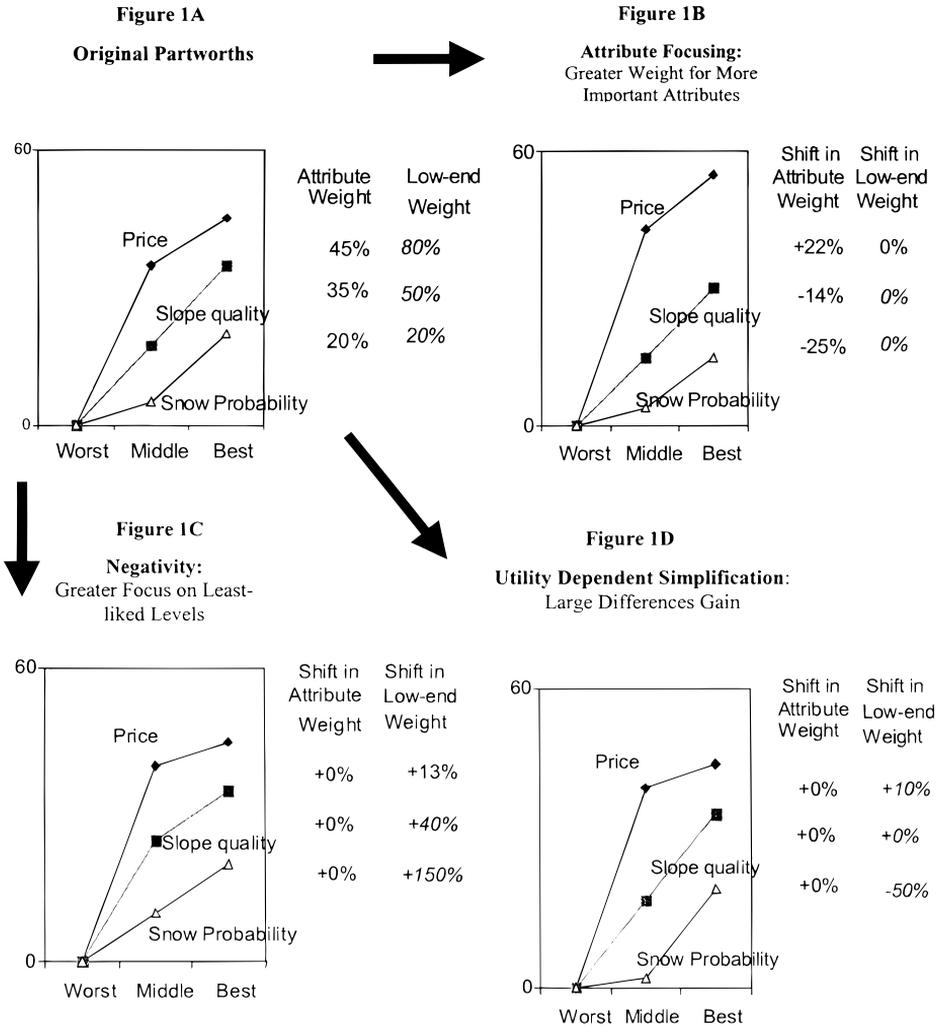
FIG. 1. Illustrating attribute focusing, negativity, and utility-dependent simplification. Attribute weights reflect the range of utility for an attribute divided by the sum of those ranges for all attributes. Low-end weight is the percentage of an attribute's utility range accounted for by the difference between the middle and worst level.

weight to attributes represented in units similar to those of the response variable (Delquié, 1993; Fischer & Hawkins, 1993; Slovic, Griffin, & Tversky, 1990; Delquié, 1997; Borcherding, Eppel, & von Winterfeldt, 1991). This distortion arises when a stimulus coded in units similar to those of the response scale is more "compatible" with that response and therefore receives greater weight. For example, an attribute with a 0–100 coding will have greater slope if the evaluation scale shares the same metric, presumably because it is easier to transfer comparable units.

Level Focusing

A second simplification mechanism involves giving exaggerated attention to particular level differences within attributes. We define level focusing in terms

of an attribute's low-end weight, the proportion of weight given to the difference between the lowest and the middle levels compared to the total utility range $[(V_{mid} - V_{low})/V_{tot}]$. Thus, the target in Fig. 1A shows *price* with 80% of its weight in the low end, demonstrating diminishing returns to better (lower) price. *Slope quality* has constant returns, so its low-end weight is 50%. Finally *probability of snow* has increasing returns evidenced by a low-end weight of 20%. We examine two mechanisms that can lead to shifts in level focusing—negativity and utility-dependent cutoff strategies. Negativity involves giving greater attention to less preferred attribute levels. The contrast between Figs. 1A and 1C illustrates this process whereby the differences between the high and middle levels diminish, and those between middle and low levels increase. In particular, the low-end weight of price increases by 13% (80 to 90%), slope quality by 40% (50 to 70%), and snow probability by 150% (20 to 50%). Negativity effects have been demonstrated in a large number of domains (Kanouse & Hansen, 1972; Wright, 1974; Taylor, 1991; Wedell & Senter, 1997). We test whether negativity also occurs in an agent task, and if its magnitude changes across the three different elicitation tasks.[1]

Reference dependence is a largely accepted theoretical driver of negativity. Following Kahneman and Tversky's (1984) prospect theory, value functions are steeper below the reference point than above it. This loss aversion around a reference point predicts negativity as long as the reference point is near the middle level of an attribute. Reference dependence should have differential impact for rating, choice, and matching. Rating tasks are likely to evoke anchoring around the middle levels of an attribute, leading to lower valuations of alternatives containing low attribute levels. For choice, this reference dependence will be further exacerbated if options are more likely to be eliminated when one or more attributes fall below minimum acceptable reference levels, producing an apparent kink in the value function at that reference point. By contrast, negativity is least likely when matching pairs since they provide their own reference, lessening the need for or availability of an external reference point.

Klein and Bither (1987) suggest a different form of level focusing. Under their *utility-dependent cutoff* mechanism, people simplify judgment tasks by selectively ignoring less valued attribute differences. This mechanism is important because its focus on large utility differences is a justifiable simplification heuristic from a cost–benefit perspective. That is, if one has to ignore differences among levels, it is most efficient to ignore small differences that will minimally impact preferences. As illustrated in the contrast between Figs. 1A and 1D, this process expands the larger value differences within an attribute and diminishes the smaller ones, thereby exaggerating any initial curvature. Klein and Bither produced evidence that cutoffs follow a utility-dependent

---

[1] Logically, positivity is also a possible screening mechanism, whereby alternatives are only evaluated further if they contain the highest level of an attribute. However, with the exception of Edland and Svenson (1993) we know of no example where positivity has been found as a screening mechanism.

model, but were not able to separate utility dependence from negativity. We develop experiments that expand their work by testing contexts in which negativity and utility dependence produce conflicting predictions.

Below we examine how these distortions can be expected to differ among three different tasks. Table 1 displays the particular tasks used: ratings of individual alternatives, choices among triples of alternatives, and matching pairs of alternatives to indifference.

Choice involves the selection of one alternative from a set, where each alternative is defined as a collection of different attribute levels. Contrasting choices

TABLE 1

Examples of Three Preference Elicitation Tasks

A. Choice

Which item would you choose?

	A	C	B
Total cost ($900/$600/$300)	$900	$600	$300
Ski Slope Quality (C-70/B-80/A-90)	90	80	70
Likelihood of Excellent Snow (50%/70%90%/%)	90%	70%	50%
Travel time (6/4/2 hours)	4 h	3 h	2 h
Night Life (poor/fair/good)	poor	good	fair

B. Rating

Rate the overall value of this ski trip

Total cost ($900/$600/$300)	$300
Ski Slope Quality (C-70/B-80/A-90)	70
Likelihood of Excellent Snow (50%/70%/90%)	50%
Travel time (6/4/2 hours)	2 h
Night Life (poor/fair/good)	fair

Worst   Average   Best

1   2   3   4   5   6   7   8   9

C. Matching

Indicate the cost that would make these two trips equally valuable

	A	B
Total cost ($900/$600/$300)	?	$300
Ski Slope Quality (C-70/B-80/A-90)	90	70
Likelihood of Excellent Snow (50%/70%/90%)	90%	50%
Travel time (6/4/2 hours)	4 h	2 h
Night Life (poor/fair/good)	poor	fair

among triples with the monadic rating and binary matching tasks, our choice task gives agents the most information to process. Further, because a respondent's goal is to select one, rather than rate or evaluate each alternative, there is value in heuristics that facilitate a reasonable decision without too much effort (Wedell & Senter, 1997). For choice, the confluence of a large amount of information with a task that encourages heuristics leads to the expectation that choice will be the most susceptible to both attribute and level simplification. Previous research leads us to predict two specific forms of simplification in choice. First, consistent with the prominence effect, we expect choice to put the greatest weight on the most important attribute. Second, with respect to level focus, we anticipate that choice will focus on negative attribute levels as respondents use the less preferred levels of attributes as a convenient way to screen out or quickly devalue alternatives.

The *rating* task, in contrast to choice or matching, focuses on individual alternatives and thereby requires the processing of the fewest pieces of information (see Table 1A). Since it generates the lowest information load, it should be the fastest and evoke the least simplification. In particular, people should be able to process more attributes, leading to less attribute focusing. Another differentiating characteristic of ratings is that they are made relative to implicit norms. That is, in choice and matching, the alternatives are directly compared with one another, while in a rating task each alternative is evaluated by itself, with the references to past alternatives largely being carried in memory. Thus, for ratings, the upper and lower bounds of the attribute levels across alternatives offer a frame of reference, while moderate attribute levels provide a natural reference point. This reference dependence combined with loss aversion leads to a prediction of a negativity bias for ratings.

Matching between pairs combines the self-anchoring qualities of choice with the relative simplicity of a rating task. Instead of focusing on the value of an alternative, attention is on the value of *differences* between alternatives. Thus in Table 1C, a person might first evaluate the value of a 20-point difference in *snow quality,* followed by a 40-percentage-point difference in the *probability of good snow.* To simplify the difficult process of valuing cross-attribute differences, we expect respondents to focus first on the salient attributes, giving them greater weight.

Another likely attribute bias for matching comes from scale compatibility. We predict that the matching attribute will receive too much emphasis. For example, if price is the matching variable, assessing the dollar value that makes the two alternatives equal in value draws attention to price relative to other attributes. Further, if the respondent *anchors* on the price given and then *insufficiently adjusts* for the other attribute differences, then the anchoring and adjustment process leads to an overestimation of the importance of the matching variable (Tversky, Sattath, & Slovic, 1988). For example, in the matching task in Table 1C, anchoring on and insufficient adjustment from the price of $300 will result in an increase in the derived value of price.

Borcherding, Eppel, and von Winterfeldt (1991) demonstrate the distorting power of scale compatibility in a matching context. They compare various

attribute importance estimates: "ratio," "trade-offs," and, "swing weights," all asking for judgments of the value differences between attributes where the matching variable rotates across the different attributes. A fourth method, "pricing-out," is similar to our matching task in that price consistently serves as both an attribute and the response scale. Borcherding, Eppel, and von Winterfeldt (1991) find that the derived importance of price is 10 times greater for pricing-out compared with the other three methods. The magnitude of this difference suggests that agents in our matching task will put too much weight on to the matching attribute.

In contrast to attribute focus, the pairwise nature of the matching task leads us to expect minimal level focus in the matching task. The "concreteness principle" asserts that "information that has to be held in memory, inferred or transformed in any but the simplest ways, will be discarded" (Slovic & MacPhillamy, 1974). Applying this principle suggests that people will tend to focus on differences (e.g., the 2-h difference between 4 and 6 h) but ignore the average level, since that takes extra work. To the extent that the information about the general level of the pair is discarded, then the matching task can be expected to show less differential level focusing compared with choice and matching. For that reason, if any bias is likely for a pair task, it is to "overlinearize" value trade-offs by establishing a constant rate of substitution between a given pair of attributes, regardless of the level of each.

In this article, we present two studies that test these expectations. In the first study, the relationship between the target attribute levels is linear—the value of going from the lowest to the middle level is equivalent to the shift from the middle to the highest level. This linear partworth study provides a test of level and attribute distortions where it is relatively easy for respondents to understand and translate the differential trade-offs between attributes. In the second study, the relationship of levels within attributes is nonlinear, sometimes increasing and other times decreasing with improvements in an attribute. This nonlinear partworth study tests the generality of our results in a more cognitively demanding context and better discriminates among rival theoretical mechanisms.

THE LINEAR PARTWORTHS STUDY

Eighty MBA students participated in a study administered entirely by personal computers. We asked respondents to imagine working for a company that selects and markets ski vacations. Bar graphs, such as those shown in Fig. 2, displayed the values for different levels of attributes of ski vacations. Respondents were then challenged to apply these values to the selection and evaluation of ski trips the company might offer. They received $10 for participating and an additional monetary reward of around $5 depending on how accurately their judgments matched the displayed values. The exercise had three parts; first, an introductory and training section; second, the actual choice, rating and matching tasks; and third, a section that assessed subjects' own attitudes toward the tasks.
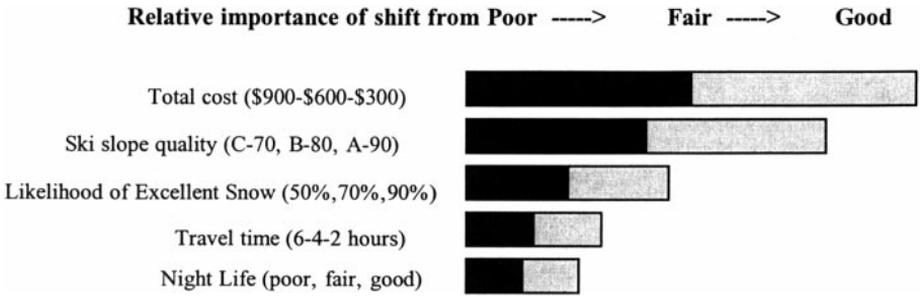
FIG. 2.   Lengths of bars for each attribute reflect the relative utility provided by each attribute. The dark part for reflects the utility value of moving from poor to fair (dark), while the lighter segment reflects the value of moving from fair to good.

Training

   To help respondents understand how to apply the company's values to decisions, they participated in training tasks involving simple choices and matching to indifference. For example, the first training task, shown in Fig. 3A, requires a choice between a $300 plan with a "poor" (70) slope quality against a $900 plan with "good" (90) slope quality. In this case, the correct choice is the inexpensive plan, since the length of the bar in Fig. 2, reflecting the $900/$300 price difference, is greater than the bar reflecting the poorgood quality difference. We congratulated those making the correct response and moved them to the next choice. An incorrect response evoked an explanation for why the low-cost alternative is preferred, saying, "the importance of $300 vs. $900 in total cost is greater than the importance of 70 vs. 90 in quality." Analogous feedback continued for the next six choice training tasks.

A   Training exercise: Which one would you choose?

	A	B
Total Cost ($900-$600-$300)	$300	$900
Ski slope quality (C-70, B-80, A-90)	70	90

B   Training exercise:
What Cost Would Make These Equally Valuable?

	A	B
Total Cost ($900-$600-$300)	$300	?
Likelihood of Excellent Snow(50%,70%,90%)	50%	90%

FIG. 3.   (A) Example of the first choice training exercise. (B) Example of the first matching training exercise.

Training with nine matching tasks followed. In these exercises, respondents estimated the level in one attribute that would make two alternatives equally valued. For example, they had to estimate the price of a plan with a 90% chance of snow that would equal a $300 plan with a 50% chance of snow (see Fig. 3B). After generating their estimates, respondents learned the correct answer ($570) and received praise appropriate to the accuracy of their responses. An answer within 10% elicited a "Very good" response; errors of 10–20% produced an "OK," and errors greater than 20% evoked, "That's not very accurate."

We designed this training program to enable respondents to associate values of attribute levels with the lengths of the lines. However, by providing neither a ruler nor numbers we intentionally made it difficult for respondents to apply a mechanical rule. Further, the subsequent tasks differed on five attributes, rather than on two as in the training tasks, requiring that subjects generalize the idea of compensatory attribute trade-offs to a far more complex task. The choice and matching tasks were designed to enable respondents to understand the meaning of relatively simple trade-offs between attributes. There were no training tasks for rating because the rating values change in complex ways as the number of attributes changes. In order to help subjects become acquainted with the rating task with five attributes, we described the best and worst alternatives and indicated that they were the best (rated as 9) and worst (rated as 1). In this way, subjects could understand both the range of products and how they mapped onto the possible responses.

Preference Elicitation Tasks

Following the training session, each subject completed 18 rating, 18 matching, and 18 choice judgments corresponding to those shown in Table 1. The rating judgments each described one alternative and asked the subject to assign a rating between 1 (*worst*) to 9 (*best*). The matching tasks each had two stages. In the first stage, a respondent chose between two alternatives defined on all attributes but price. In the second stage, the computer defined the price of the less preferred alternative and asked the price of the preferred one for them to be equally valued. Finally, the choice tasks required a simple selection of the best from three alternatives. While performing these tasks the partworths shown in Fig. 2 were always in view. Across respondents, we randomized the order of the three tasks.

We generated stimuli using related, but differing methods. The rating task came from an 18×5 orthogonal array (Addelman, 1962) which permits all main effects for the five attributes each at three levels to be estimated with maximum efficiency. For the matching task, we built a pair design from the same array with the following recoding: we replaced all level 1's with a pair having level 1 on the left and 2 on the right, all level 2's with a 2 on the left and 3 on the right, and all level 3's with 3 on the left and a 1 on the right. Finally, for the choice task, we used the following cyclic rule to generate choices: an attribute with level 1 generated three choices with levels 1, 2, 3; that with level 2 generated choices with levels 2, 3, 1; and level 3 translated into a 3, 1, 2.

An additional aspect of this study investigated whether different attribute labels would affect the results. As Table 2 shows, the 80 respondents were randomly assigned to one of four conditions with different labels attached to the first- and second-most important attributes. Condition 1 reflects the labels shown in Fig. 2, with five attributes, in order of importance, being *total cost*, *slope quality, probability of good snow, travel time*, and *night life*. In condition 2, *total cost* changes position with slope quality. Similarly, in conditions 3 and 4, *waiting time at the lift* replaces *total cost* in conditions 1 and 2. Across labeling conditions, the target partworth utilities stayed the same, only the labels changed. Matching was always done in terms of the first (most important) attribute. Somewhat to our surprise, we found that the derived partworths and accuracy differed little despite these substantial labeling differences. Subjects were able to learn the appropriate trade-offs despite heterogeneous prior orientation to the labels. Thus, for our purposes here, we treat the labeling conditions as four independent replications of the experiment. To the extent that the results hold across these different labeling conditions, we can feel confident that they hold generally.

To assess consistent biases among the methods, we estimate coefficients within each of the tasks from data pooled across respondents. These coefficients estimate the values that respondents actually applied within each of the tasks. Biases can be estimated by comparing the derived and target (true) partworths. For the ratings task, a dummy-variable regression estimated an additive model that best predicted these ratings. For matching, a similar regression on level differences (e.g., the difference between high and low snow quality) predicted the value of the differences of the matching variable. Finally, for choice, multinomial logit (Maddala, 1983) produced analogous coefficients that maximized the likelihood of the choices made.

The resulting scales then differ with respect to the zero points for each attribute and their general metric. Adding a different constant for each attribute makes no difference for predicting choices since those constants are added to each alternative. Thus for display purposes the lowest (least preferred level) of each attribute is set to zero. Then to put the outputs from the three tasks

TABLE 2

Four Labeling Conditions in the Linear Study

Condition weight	Attr. 1 36%	Attr. 2 28%	Attr. 3 16%	Attr. 4 11%	Attr. 5 9%
1	$300–$900	90–70 slope	90–50% Snow Probability	2–6 h	Good/poor night life
2	90–70 slope	$300–$900	90–50% Snow Probability	2–6 h	Good/poor night life
3	15 to 45-min wait	90–70 slope	90–50% Snow Probability	2–6 h	Good/poor night life
4	90–70 slope	15 to 45-min wait	90–50% Snow Probability	2–6 h	Good/poor night life

in the same metric, each is multiplied by a positive constant that best reproduces the target partworths. This affine transformation was determined by a simple regression through the origin of the true against predicted partworths. These transformations of origin and scale permit a focus on the relative partworths in such a way that preserves the rank order of partworths. More important, the transformations do not affect our two critical measures, attribute importance and low-end weight.

Results from the Linear Study

Figure 4 presents the partworths for the three methods against the target values and Table 3 summarizes the biases of the three tasks with respect to attribute and level focus biases, decision time, and attitudes. The tests of significance use the four labeling conditions and three tasks as factors in a two-way ANOVA. Throughout, the contrasts between the four labeling conditions are not significant ($p > .10$) and are not discussed further.
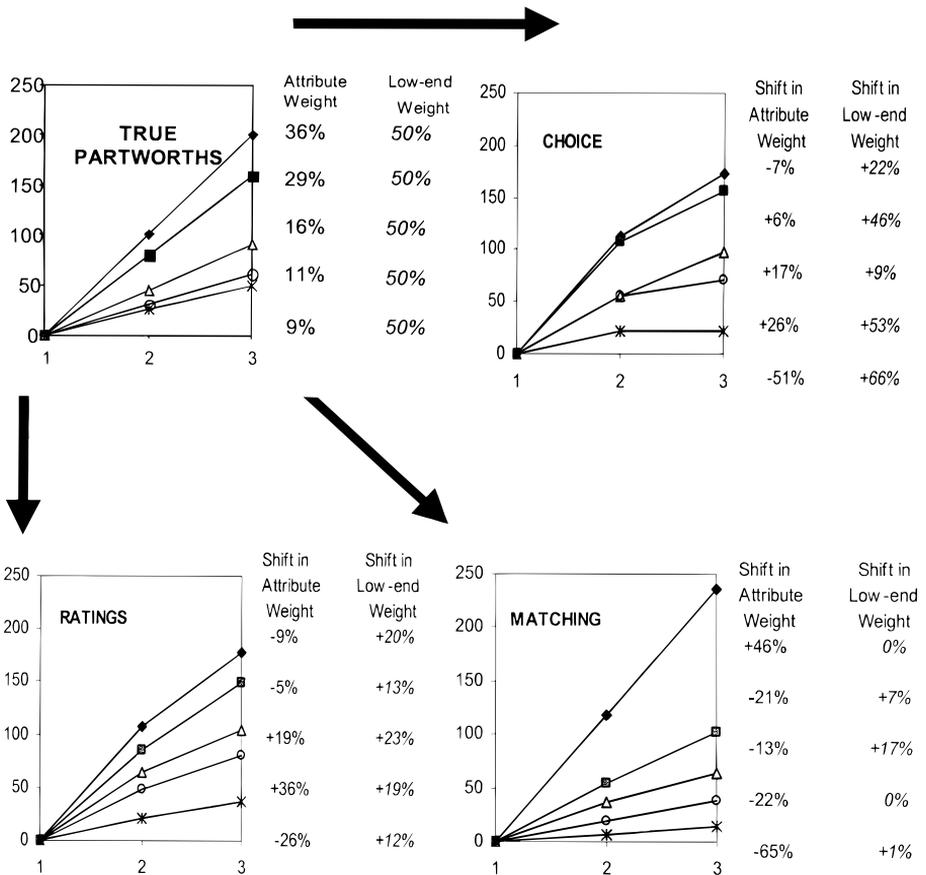

FIG. 4.   Partworth values for linear partworths study for choice, ratings, and matching. Also shown are the percentage shift in attribute weight and low-end weight compared with the target partworths shown.

TABLE 3

Results of the Linear Trade-Off Study

	Choice	Ratings	Matching
Attribute Focus			
Percentage that the top (matching) attribute is overweighted	$- 7\%_b$	$- 9\%_b$	$46\%_a$
Level Focus			
Percentage increase in low-end weight	$39\%_a$	$18\%_b$	$6\%_c$
Decision Time			
Time per judgment (in seconds)	19_b	11_c	26_a
Attitudes (0–100)			
Realistic	67_a	61_b	53_c
Confident	63_a	57_b	43_c
Easy	58_a	49_b	39_c
Interesting	57	56	55

Note. Items sharing different subscripts are significantly different $p < .05$.

Consider first shifts in attribute focus for the most important attribute displayed in Fig. 4.[2] For choice, the most important attribute drops in weight by 7%, while attributes with moderate importance gain. Ratings present the same pattern, with a 9% drop in the importance of the most important attribute. By contrast, matching displays a very different pattern, with the most-important attribute increasing by a striking 46%. The drop in weight for the most important attribute is not significant for choices or ratings, in contrast to a significant positive gain for matching. Thus, these results provide no evidence for a prominence effect in choice but substantial evidence for a compatibility effect in matching.

Looking for biases within attributes, Fig. 4 demonstrates consistent shifts in low-end weight for choice and ratings. This negativity is visually apparent for choice and ratings by the downward curvature indicated, but is hard to detect visually in the case of matching. Indeed as Table 4 indicates, choice overweights the low-end levels by an average of 39%, while ratings overweight them by 18% and matching by only 6%. The biases for both choice and ratings are significantly greater than zero ($p < .05$), while that for matching is not significant ($p > .10$). Thus, as predicted, choices, and to a lesser extent ratings put unjustified emphasis on negative information, while the matching task, with its focus on differences between attribute levels, appears less affected by this bias.

Finally, we note the time taken and attitudes toward the tasks. Rating is fastest, consuming an average of 11 s for each of the 18 judgments. Choice among triples is next at around 19 s, followed by matching at 26 s. One of the reasons matching takes so long is that it involves two separate tasks; the initial

---

[2] There is a large, 51%, loss in importance weight for the least important attribute. However, its absolute loss compared with the magnitudes of changes of the more important attributes is relatively small.

choice among a stimulus pair averages 12 s, and then matching to indifference takes another 14 s. The three tasks also differ with respect to respondent attitudes. Choice is rated easiest; respondents are more confident that they are correct, and the task is seen as most realistic. Ratings are in the middle, and matching performs least well on these perceptions of ease, confidence, and realism.

Discussion

The results from the first study were quite surprising. The prominence effect suggested that choice would put too much weight on the prominent attribute (relative to the target value), whereas the compatibility effect would put too much weight on the matching attribute, which in our design was also the most important attribute. Extrapolating from past findings, we had expected the prominence bias to be the larger of the two, leading to greater overweighting of the most prominent attribute in choice compared to matching. Instead, we found the opposite—a slight underweighting of the most important attribute for choice and rating along with a substantial overweighting for the matching task.

In addition, we find a negativity bias of nearly 40% in choice and nearly 20% in ratings. These differences are large enough to affect the rank ordering of the partworths. If we rank order the expressed partworths for choice and rankings, we find that the low-end partworths have consistently higher rank importance compared with those reflecting the high end. Furthermore, since no significant negativity bias is apparent in the matching judgments, it is unlikely that the negativity bias for choice and ratings could have arisen from an internal reevaluation of the input data. Instead, the negativity bias appears to reflect the ways the given values are *expressed* in the tasks. In choices and ratings, people act as if they automatically treat differences on the low end of each attribute as mattering more than comparable differences on the high end, whereas in matching the value difference is quite independent of the level.

This lack of a negativity bias in matching may be due to two factors. First, since the target partworths were linear, matching may simply be better at approximating these true values. Alternatively, by focusing on differences, matching may be biased toward the linear, equal spacing of level differences regardless of the true level differences. To discriminate between these two accounts, we designed a second study with attributes whose target partworths either displayed negativity (decreasing returns to a fixed improvement in the variable) or positivity (increasing returns). If matching is biased toward producing equally spaced partworths, this bias should be apparent in these nonlinear conditions.

Having target attributes whose partworths show both increasing and decreasing returns offers a further theoretical advantage. It enables us to distinguish between a simple negativity bias and the Klein and Bither (1987) utility-dependent cutoff mechanism. Under negativity, the lowest levels of an attribute will increase in importance. However, under a utility-dependent model, only the larger utility differences (whether between positive or negative levels of

an attribute) will be inflated. Thus, attributes with increasing returns (such as *snow probability* in Fig. 1A), should see greater curvature if utility-dependent focusing is correct (Fig. 1D), but should see that upward curvature moderated if negativity is more salient (Fig. 1C).

NONLINEAR TRADE-OFFS STUDY

The experimental procedure was similar to that of the linear study except for three changes. First, rather than rotate labels, all subjects experienced the labeling condition that had price as the most important attribute. Second, two different curvatures of target partworths were manipulated between participants. Third, since the utility structure underlying this curvature was more complex, the training expanded from 7 to 11 choices and from 9 to 12 matching tasks.

Sixty MBAs were randomly assigned to one of two conditions. Condition 1 placed greater weight on the negative levels of the second and fifth attribute and less weight on the negative levels of the third and fourth attribute, as shown in Fig. 5A. Condition 2, shown in Fig. 5B, reversed the curvature of

FIGURE 5A
Values for the Nonlinear Study--Condition 1
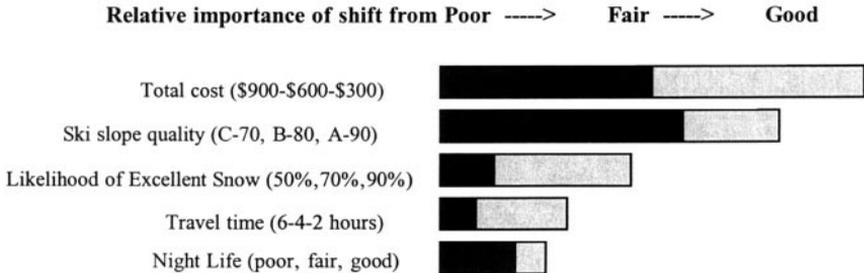
FIGURE 5B
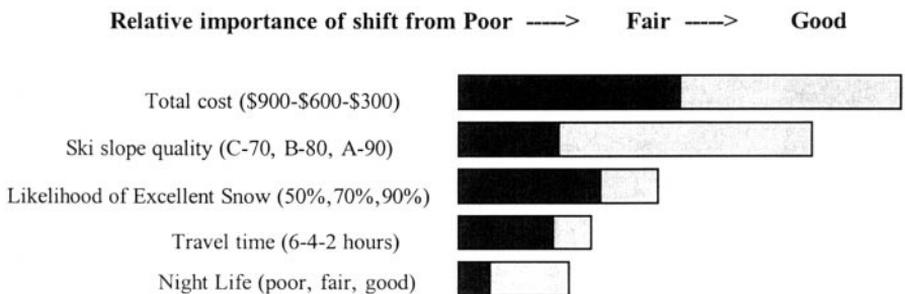**Values for the Nonlinear Study--Condition 2**

FIG. 5. (A) Values for the nonlinear study—Condition 1. (B) Values for the nonlinear study—Condition 2.

condition 1 for each attribute, except the first, which was linear in both conditions. The conditions were designed so that the average of the two conditions was equivalent to the linear partworths in the earlier study.

Results from the Nonlinear Study

Table 4 summarizes the bias and attitude statistics, while Fig. 6 graphs the derived partworths averaged across the two initial curvature conditions. These results are remarkably parallel to those in the linear study. We consider first biases in attribute focus, followed by those related to level focus.

In terms of attribute focus, Table 4 shows that choice and ratings again give less weight to the first attribute than is appropriate, while matching again gives it more. Both choice and ratings display an attribute focus bias that is in a direction opposite to that of a prominence effect, but not significantly so. In contrast, the matching task displays a strong and significant scale compatibility bias that overvalues the matching attribute. This 23% overvaluation of the matching variable may be substantially less than the 46% found in linear study, but still remains a substantial problem for the matching task.

Thus far we have emphasized the weight given to the most important attribute. However, given the unanticipated lack of a prominence effect in the choice data, it is appropriate to examine the weights given to the less important attributes as well. Defining attribute weight as the utility range for each attribute divided by the sum of those ranges, Fig. 7 graphs target importance weights against expressed attribute importance weights for the three tasks. The diagonal shows where weights would be if they were perfectly expressed.

Both panels in Fig. 7 display the above-mentioned overweighting of the matching variable and a somewhat smaller underweighting for choice and ratings. The new insight from these graphs is that for both choice and ratings the position of the middle attributes above the diagonal indicates that they

TABLE 4
Results of the Nonlinear Study

	Choice	Ratings	Matching
Attribute Focus			
Percentage that the top (matching) attribute is overweighted	$-9\%_b$	$-21\%_b$	$23\%_a$
Level Focus			
Percentage increase in low-end weight	$30\%_a$	$21\%_a$	$-5\%_b$
Decision Time			
Time per judgment (in seconds)	48_a	20_b	53_a
Attitudes (0–100)			
Realistic	74_a	51_b	52_b
Confident	59_a	51_b	51_b
Easy	50_a	47_a	37_b
Interesting	54_b	52_b	59_a

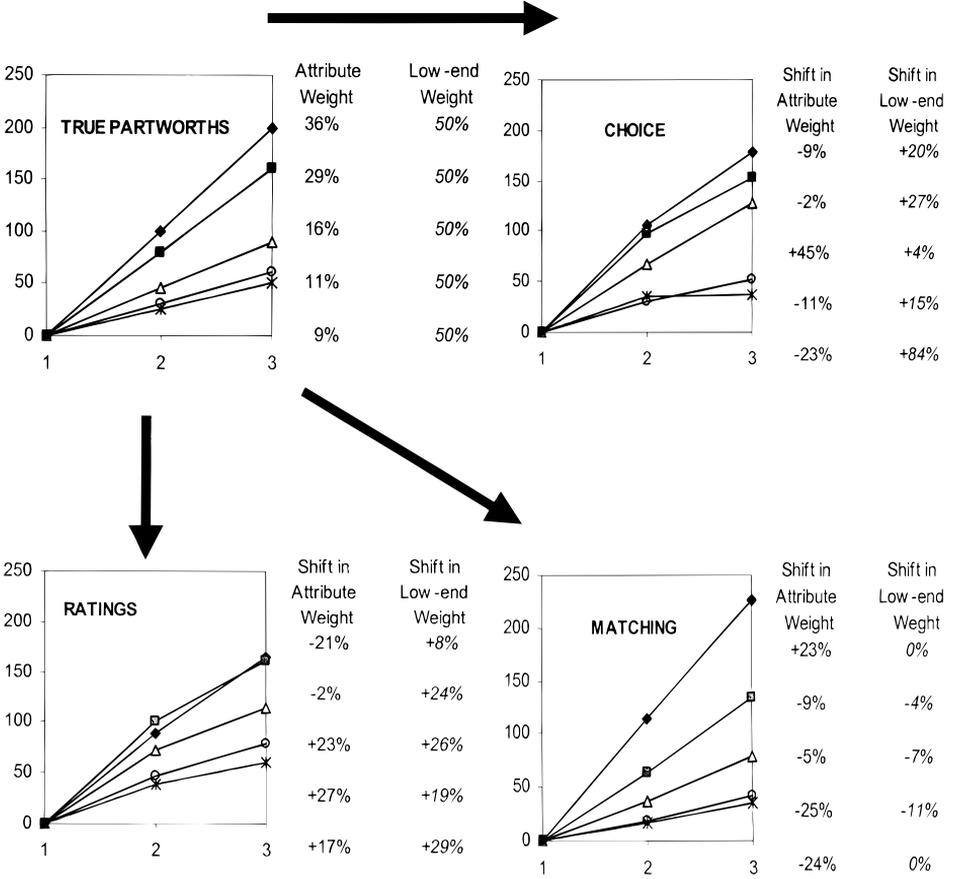*Note.* Items with different subscripts are statistically different, $p < .05$.

FIG. 6. Nonlinear partworths averaged across both curvature conditions. Also shown are the percentage shift in attribute weight and low-end weight compared with the target partworths.

are given more weight than is appropriate. An equivalent way to interpret this result is that the three most important attributes are weighted more equally than is justified. That is, if we compute the slopes between the expressed and the target weights for the three most important attributes, they average $m = .68$ for the linear and $m = .57$ for the nonlinear study. Both are significantly ($p < .05$) lower than the 1.0 they would be if attribute weights were correctly expressed. This equal-weight bias could be driven by the fact that the true values for each attribute are prominently displayed in our agent task, making it less reasonable to focus on just one attribute and encouraging simplification by giving equal weight to each attribute considered. The equal weight bias generalizes a result found by Russo and Dosher (1983). They found an equal weighting bias in paired comparisons, whereas we show it also occurs in choices among triples and for monadic ratings.

Turning attention to level focus, Fig. 6 shows that choices and ratings again produce a visually apparent negativity bias. Table 4 shows that low-end weight increases an average of 30% for choices and 21% for ratings, in contrast to a nonsignificant 5% decrease for matching. Thus, the nonlinear study replicates
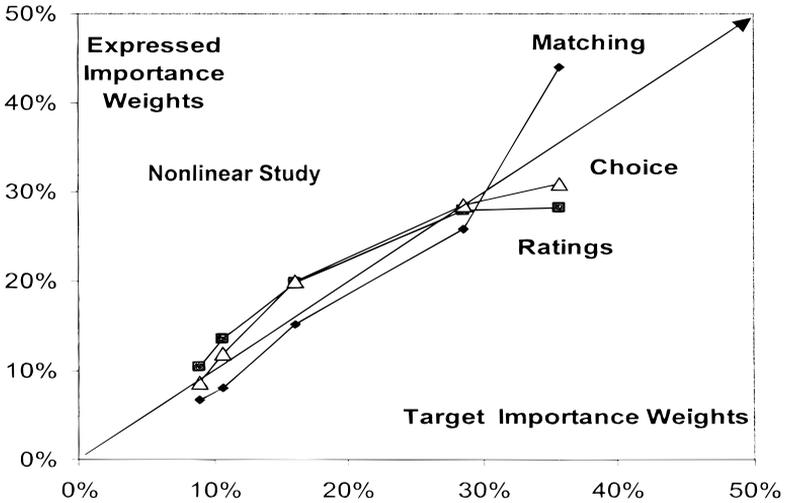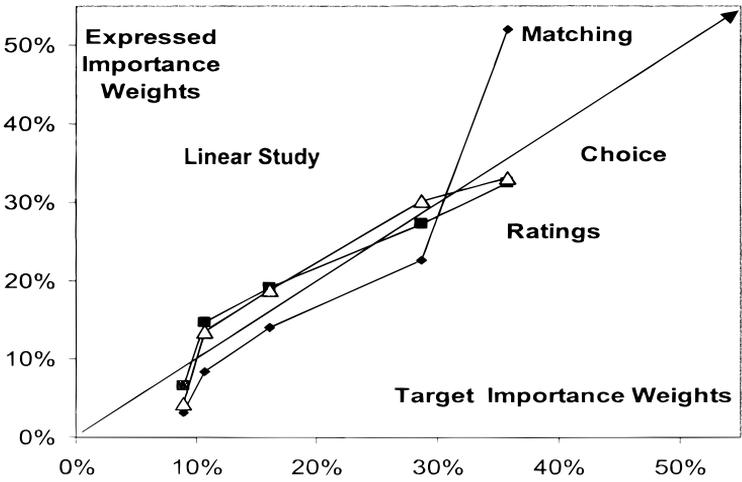
FIG. 7. Expressed vs target attribute importance weights. Importance weights are the ratio of the range of utility for each attribute divided by the sum across all attributes.

the linear study, showing that choice and ratings produce significant negativity, while matching does not.

The partworths shown in Fig. 6 are appropriate for estimating the general tendency to give too much weight to low-end levels, but it is important to recall that they reflect averages across target conditions that differ in their low-end levels. Figure 8 groups attributes that share the same target curvature. The left panel shows attributes with positive targets, reflecting *likelihood of excellent snow* and *travel* time in condition 1 and *ski slope* and *night life* in condition 2 (see Fig. 5). The right panel displays the curvature for these same attributes with negative target curvature.

The contrast in expressed curvature between positive and negative target

attributes is important because it permits a test of negativity against utility-dependent distortions (Klein and Bither, 1987). Recall that under utility dependence large *differences* are given greater weight, while smaller ones are given less weight. In terms of weight given to low-end levels, utility dependence predicts that the small low-end weight of the positive target will become even smaller. By contrast, negativity predicts an increase in low-end weights in all conditions. If both processes operate, we would expect to see more moderate biases for the positive targets because utility dependence would cancel negativity, but greater biases given negative targets because both processes operate to increase negativity.

As Fig. 8 shows, the reverse occurs. With positive attributes that place minimal weight on the low end (29%), both choice and rating display expected negative bias. By contrast, when the target already has strong negativity (72%), these biases are moderated or even reversed. Put differently, the general negativity bias for choices and ratings comes largely from the condition in which the initial target has increasing returns, a result that offers virtually no support for the utility-dependence model.

Figure 8 is also useful in contrasting the sensitivity of the three tasks to different target conditions. For example, matching is quite accurate in tracking the correct curvature, while choice appears to display consistent negativity. Ratings, by contrast, show the least impact from target curvature. The curvature expressed by ratings is muted, differing very little across the two conditions.

Finally, Table 4 gives other measures of differences between the tasks. For the more complex study, decision time increased by 29 s for choice (19 → 48 s) and by 28 s for matching (25 → 53 s) but by only 9 s for the already fast ratings (11 → 20 s). These differences suggest that extra time may be more valuable in choice and matching compared to ratings. In choice and matching
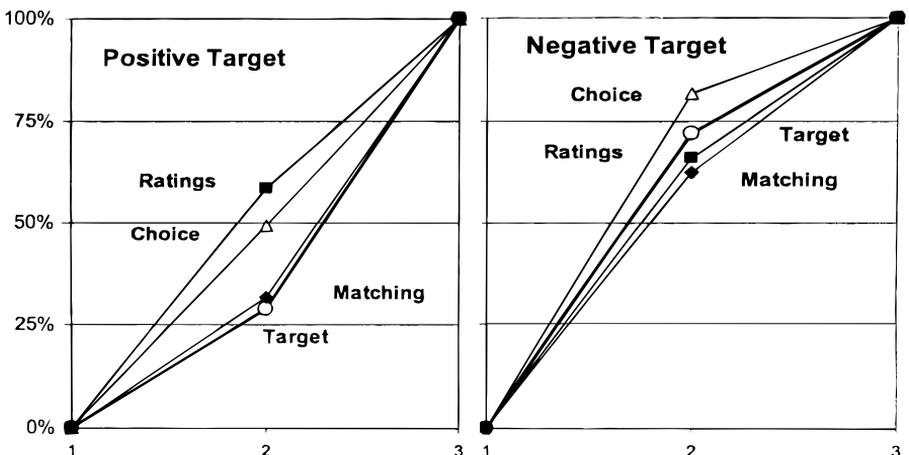

FIG. 8.   Average low-end weight conditioned on the curvature of the target. Low-end weight is the utility range of an attribute accounted for by the difference between the middle and the worst levels. A positive target implies increasing returns to improvements (left panel) while a negative target (right) implies decreasing returns.

it is possible to know what makes a good decision, whereas for ratings, additional effort may not be expended due to uncertainty projecting values onto an arbitrary rating scale. In terms of attitude toward the task, choice still dominates in being perceived as the most realistic and remains the easiest task about which respondents feel most confident, but matching now surpasses it in terms of being perceived as the most interesting.

DISCUSSION AND CONCLUSIONS

The purpose of this article has been to examine the impact of task on the degree to which agents can consistently express the known values of the principal whose interests they represent. Such agent tasks are important not only because they allow us to have veridical measurements of judgment accuracy, but also because there are many contexts in which decision-makers express values of others through choices, ratings, or matching judgments. While our tasks are admittedly simple and somewhat stylized, the biases evident in these simple cases could portend even greater biases in cases where policies are not as well defined. After all, our experiments minimize distortions from understanding or learning values and focus on distortions arising from the final expression of values in choices, ratings, and matching.

Possible Alternate Explanations

Before examining the implications of these results, it is important to consider whether they could have been generated by other mechanisms. Specifically, it is important to consider whether they could have arisen either through a rank order transformation of the original partworths or through noisiness in our subjects' responses. The rank order explanation assumes that respondents encode only the rank order information from the original partworths bar graphs. Under this assumption the expressed and true partworths should then be related only by their rank orders. However, an examination of the rank order of the target against expressed partworths reveals consistent, rather than random, deviations from the initial rank orders. In particular, expressed orderings of partworths for choices and rankings favor negativity in more than 80% of the cases, a result incompatible with an account based on a rank order transformation of the original partworths.

A second hypothesis that initially seemed feasible is that these results could have been an outcome of noise (variability) either within or between subjects. To test that possibility we ran a series of analyses simulating responses with different levels of noise. In the noisy conditions, differences between the partworths became muted and less consistent, but we were unable with noise either to produce negativity or the pattern of attribute weights we found. Of course, differential variability applied to different attributes or attribute levels could produce our results. For example, we could simulate negativity by injecting greater precision into the evaluation of the low-end attributes. However, it

would be difficult, and probably not productive, to pit such a noise-based model against simply applying greater weight to these attributes.

Summary of Results

Choice.   Given the noncompensatory heuristics that have been associated with choice, the important finding from our study is that people can learn to make choices that do a good job of trading off attributes. We were surprised that agents showed only minimal levels of attribute simplification. It appears that when policies are clearly articulated, motivated agents make choices that reflect compensatory trade-offs among attributes. The primary attribute focus error appears as a bias toward equal weighting of the top three or four attributes.

The weak point for choice is a consistent overvaluation of negative information. This result is somewhat surprising because an effort–accuracy framework predicts that large utility differences should generate greater attention regardless of valence, but it follows from well-known process and perceptual accounts. The process account posits that alternatives are initially screened for negative values (Wedell & Senter, 1998; Russo & Leclerc, 1994). In that case, the negative values get more weight because the positive values never have a chance to counterbalance them. The perceptual account posits that a strong predisposition toward loss aversion overrides the training information. It is quite likely that both the process and the perceptual distortions jointly contribute to the observed overweighting of negative attribute levels in choice. Whatever its source, the strong negativity bias found suggests that agents may require special training to limit such loss-averse behavior when making choices on behalf of their principals.

Ratings.   Ratings present an interesting paradox. They take less than half the time of the other tasks and track target values adequately. However, they are not precise and, given respondents' expressed lack of confidence in ratings, are not perceived to be so. Part of the problem may be due to a lack of explicit training for ratings. However, in our view, the more likely culprit is the difficulty in precisely translating values to a rating scale. Consequently, the ratings tend to lack crispness—both curvature and attribute differences are muted. The place for rating, then, is as a quick, easy task to express roughly held preferences. Under such conditions, ratings are best at expressing values that are moderate in the sense of being loss averse and not too complex.

Matching.   Matching is the most difficult task for subjects and the one that consumes the most time. It is also the one that was best able to match curvature. The important caveat, however, is that matching consistently produces a substantial upward bias in the utility attached to the matching variable. This overweighting of the matching variable suggests, for example, that asking employees to attach a dollar value to a new health care benefit will lead to an underestimation of the actual value because the matching variable (money) receives too much weight. However, apart from the matching variable, the

matching task tracks true attribute weights extremely well, displaying only a minor, if consistent, diminution of curvature. Thus, asking employees decision makers to attach a dollar value to health care benefits X and Y should lead to accurate assessments of the *relative* values of benefits X and Y, even as it underestimates the dollar value of each.

To summarize, the large and consistent differences across tasks suggests that agents either need to adjust their training to account for these biases or to frame problems so that such biases are minimized. This research only begins to scratch the surface of the behavioral issues that arise in coordinating the preferences of boundedly rational principals and boundedly rational agents. More work is needed to determine the theoretical differences that lead to attribute and level focus biases.

Attribute focus biases. Most of our results for agents are quite consistent with choices people make for themselves. The exception is that we found no evidence that agents overweight the prominent attribute in choice relative to the target weight. Indeed, for choice and especially for rating, there appears a weak but consistent *underweighting* of the most important attribute. The difference between our results and research on the prominence effect (Tversky *et al.*, 1988; Fischer & Hawkins, 1993; Fischer, Carmon, Ariely, & Zauberman, 1999) certainly stems from three procedural differences. First, our task involved five, compared with the two, attributes typical in most studies of the prominence effect. Second, our study investigated how people act as agents to express (known) preferences on behalf of another, whereas respondents in previous studies of the prominence effect constructed their own preferences. Finally, the presence of visible partworth graphs may have encouraged respondents to sequentially process several of attributes instead of simplifying to the most prominent. Given that multiple attributes were processed, a logical effort-reduction strategy is to ignore differences in their weights (Russo & Dosher, 1983).

A very different theoretical picture drives attribute focus in the matching task, but one also in keeping with previous research (Borcherding, Eppel, & vonWinterfelt, 1991). The observed overweighting of the matching variable is consistent with the scale compatibility hypothesis, in which sensitivity to a cue is greater if the response is in the units used to represent an attribute (Tversky *et al.*, 1988, Fischer & Hawkins, 1993). A likely mechanism accounting for such scale compatibility biases is anchoring and (insufficient) adjustment (Tversky *et al.*, 1988). Because the default anchor is that the two alternatives have the same value on the matching attribute, it is reasonable that people adjust the matching variable insufficiently to reflect the sum of the differences of the other attributes.

Summarizing attribute focus biases, we find strong evidence of a scale compatibility bias in matching, a result consistent with previous research on people making judgments for themselves. For choice and ratings, we show that classic prominence does not hold. When evaluating alternatives with numerous attributes, it appears that a subset of the more important attributes become equally

salient, generating a subsequent bias that weights those attributes more equally than is justified.

Level focus biases.   Examining level biases within attributes, we find strong evidence of negativity, the tendency to focus on the least preferred levels of attributes. These results correspond most closely to early work by Wright (1974) and Einhorn (1971). In our studies, negativity appears to be quite pervasive in choice, somewhat less so in ratings, and negligible in matching. Our results do not support the utility-dependent cutoffs hypothesized by Klein and Bither (1987). Indeed, finding that the strongest negatively bias occurs when the target has increasing returns directly contradicts predictions from the utility-dependent model. Instead, the negativity found is consistent with a reference point in the middle of the attribute range, leading to categorization of the lower end levels as losses.

This finding of negativity in choice and ratings is important because, in contrast to most demonstrations of negativity, our agency task allows us to show that people do *overweight* negative outcomes relative to an absolute target. In many settings, one cannot tell whether loss aversion is a bias or merely a reflection of the fact that losses have more emotional impact than gains of equal magnitude. In our choice and rating tasks, however, we found clear evidence that agents motivated to accurately represent the preferences of others gave more weight to negative outcomes than is appropriate.

From a perspective of helping people make better judgments, our results suggest that, with practice, people can learn to make choice, rating, and matching judgments that, on average, closely track desired compensatory behavior. While there were consistent biases associated with each of the techniques, we were struck by the depth of processing and consistency in the judgments of our respondents. By contrast, individual-level ratings or choices based on people's *own* values typically display more attribute simplification than we observed in the agent task. The intriguing prescriptive question is whether the provision of similar training and overt display of values that we gave our agents could help people make better choices when they are choosing for themselves.

More theoretically, it is useful to split the decision-making process into two stages: (1) creating an internal representation of the information and (2) expressing these representations through a specific task. Given these two stages, our agent-based procedure has permitted us to isolate the specific task-based biases that contaminate this second stage. A productive area for future research will be the specification of biases specific to the first stage, those that distort the internal representation of information leading to preferences.

REFERENCES

Addelman, S. (1962). Orthogonal main-effects plans for asymmetrical factorial experiments. *Technometrics*, 4, 21–46.

Ahlbrecht, M., & Weber, M. (1997). An empirical study on intertemporal decision making under risk. *Management Science*, 43(6), 813–825.

Bazerman, M., Loewenstein, G., & White, S. (1992). Reversals of preference in allocation decisions:

Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 37, 220–240.

Bettman, J. R., Johnson, E. J., & Payne, J. W. (1990). A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes*, 45, 111–139.

Borcherding, K., Eppel, T., & von Winterfeldt, D. (1991). Comparison of weighting judgment in multiattribute utility measurement. *Management Science*, 37(12), 1603–1619.

Delquié, P. (1993). Inconsistent tradeoffs between attributes: New evidence in preference assessment biases, *Management Science*, 39, 1382–1395.

Delquié, P. (1997). 'Bi-Matching': A new preference assessment method to reduce compatibility effects. *Management Science*, 43(5), 640–658.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure: studies and findings. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human decision making*. New York: Plenum Press.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance* 6, 1–27.

Fischer, G., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 580–597.

Fischer, G., Carmon, Z., Ariely, D., & Zauberman, G. (1999). Goal-based construction of preferences: Task goals and the prominence effect. *Management Science*, 45, 105–1075.

Hammond, K. R., Summers, D. A., & Deane, D. H. (1973). Negative effects of outcome feedback on multiple cue probability learning. *Organizational Behavior and Human Performance*, February, 30–34.

Hawkins, S. A. (1994). Information processing strategies in riskless preference reversals: The prominence effect. *Organizational Behavior and Human Decision Processes*, 59, 1–26.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67, 247–257.

Kahneman, D., & Tversky, A.(1984). Choices, values, and frames. *American Psychologist*, 39, 341–350.

Kanouse, D. E., & Hanson, L. R., Jr. (1972). Negativity in evaluations. In E. E. Jones *et al.* (Eds.), *Attribution, perceiving the causes of behavior*. Morristown NJ: General Learning Press.

Klein, N. M., & Bither, S. (1987). An investigation of utility directed cutoff selection, *Journal of Consumer Research*, 4, 240–256.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometric*. Cambridge, UK: Cambridge Univ. Press.

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research*, 36, 205–218.

Ordóñez, L. D., Mellers, B. A., Chang, S., & Roberts, J. (1995). Are preference reversals reduced when made explicit? *Journal of Behavioral Decision Making*, 8, 265–277.

Payne, J. (1982). Contingent decision behavior: A review and discussion of issues. *Psychological Bulletin*, 73, 221–230.

Russo, J. E., & Dosher, B. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 9, 676–696.

Russo, J. E., & Leclerc, F. (1994). An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, 21, 275–290.

Schkade, D. A., & Johnson, E. J. (1989). Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes*, 44, 203–231.

Slovic, P. (1995). The construction of preference. *American Psychologist*, 50, 364–371.

Slovic, P., & MacPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative measurement. *Organizational Behavior and Human Performance*, 11, 172–194.

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

Stone, D. N., & Kadous, C. (1997). The joint effects of task-related negative affect and task difficulty in multiattribute choice. *Organizational Behavior and Human Decision Processes*, 70(1), 159–174.

Taylor, S. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin*, 110, 67–85.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review* 95, 371–384.

Wedell, D. H., & Senter, S. M. (1997). Looking and weighing in judgment and choice. *Journal of Experimental Psychology: Learning Memory and Cognition*, 9(4), 676–696.

Wright, P. (1974). The harassed decision maker: Time pressure, distraction and the use of evidence. *Journal of Applied Psychology* 59, 556–561.