ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 67, No. 3, September, pp. 247–257, 1996
ARTICLE NO. 0077

The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives

CHRISTOPHER K. HSEE

Center for Decision Research, Graduate School of Business, The University of Chicago

This research investigates a particular type of preference reversal (PR), existing between joint evaluation, where two stimulus options are evaluated side by side simultaneously, and separate evaluation, where these options are evaluated separately. I first examine how this PR differs from other types of PRs and review studies demonstrating this PR. I then propose an explanation, called the *evaluability hypothesis,* and report experiments that tested this hypothesis. According to this hypothesis, PRs between joint and separate evaluations occur because one of the attributes involved in the options is hard to evaluate independently and another attribute is relatively easy to evaluate independently. I conclude by discussing prescriptive implications of this research.

Normative decision theories assume that people have stable and consistent preferences regardless of how the preferences are elicited. An increasing amount of evidence has appeared suggesting otherwise; for example, people may exhibit different or even reverse preferences for the same options in two normatively equivalent evaluation conditions. Most preference reversals (PRs) documented in the literature are between choice and judgment. One widely studied type of choice-judgment PRs (preference reversals) is between choice and pricing (e.g., Grether & Plott, 1979; Lichtenstein & Slovic, 1971; Slovic & Lichtenstein, 1969). In choice, participants choose between two alternatives, typically a high-payoff/low-probability gamble and a low-payoff/high-probability gamble. In pricing, participants indicate their minimum selling price for each gamble. Another widely-studied type of choice-judgment PR (preference reversal) is between choice and matching (e.g., Tversky, Sattath, & Slovic, 1988). In choice, participants choose between two alternatives. In matching, participants are presented with the same alternatives but some information is missing and participants' task is to fill in that missing information so that the two options are equally attractive.

In both the choice-pricing and the choice-matching paradigms, reversals occur between tasks that involve *different evaluation scales* (Bazerman, Loewenstein, & White, 1992; Goldstein and Einhorn, 1987). In the choice-pricing paradigm, the evaluation scale for choice is relative acceptability and that for pricing is money. In the choice-matching paradigm, the evaluation scale for choice is, again, relative acceptability, and that for matching is probability or value estimation.

The present research investigates a different type of PR than the conventionally studied choice-judgment PRs. It is between tasks that have identical (or similar) evaluation scales but different evaluation modes. Evaluation mode refers to whether the stimulus options are presented side by side and evaluated by the same people (the *joint evaluation* mode), or presented separately and evaluated by two different groups of people (the *separate evaluation* mode) (cf., Goldstein & Einhorn, 1987). In this article, I first review studies that demonstrate this type of PR. Next I propose an explanation, and describe several other studies that tested this explanation. Finally I discuss prescriptive implications of this research.

STUDY 1

Method

The purpose of this study is to illustrate the joint-separate evaluation PR effect. The study involved the

This research is supported by a fourth quarter funding provided by the Graduate School of Business, University of Chicago. Correspondence and reprint requests should be addressed to Christopher K. Hsee, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, IL 60637. E-mail: christopher.hsee@gsb.uchicago.edu. I thank Sally Blount, David Budescu, Bill Goldstein, Josh Klayman, Rick Larrick, George Loewenstein, Paul Slovic, Dick Thaler, and Elke Weber for their helpful comments on drafts of this article.

0749-5978/96 $18.00

evaluations of two hypothetical second-hand music dictionaries:

	Dictionary A	Dictionary B
Year of publication:	1993	1993
Number of entries:	10,000	20,000
Any defects?	No, it's like new.	Yes, the cover is torn; otherwise it's like new.

The questionnaire for this study had three between-subject versions, joint-evaluation, separate-evaluation-A, and separate-evaluation-B. In each version, participants were asked to assume that they were a music major and that they were looking for a music dictionary in a used book store and planned to spend between $10 and $50. In the joint-evaluation condition, participants were told that there were two music dictionaries in the store. They were then presented with the information about both dictionaries (as listed above) and asked how much they were willing to pay for each dictionary. In each of the separate-evaluation conditions, participants were told that there was only one music dictionary in the store; they were presented with the information on one of the dictionaries and asked how much they were willing to pay. (Because there was only one dictionary in each separate-evaluation condition, the label "A" or "B" was not used.) Note that across the three conditions the evaluation scale was held constant, namely, willingness-to-pay (WTP) price.

Respondents were 116 unpaid college students from the University of Chicago and the University of Illinois at Chicago. They randomly received one of the three versions of the questionnaire and completed it individually.

Results and Discussion

The results are summarized in Fig. 1,[1] As the figure shows, there was a PR between joint and separate evaluations. In joint evaluation, willingness-to-pay (WTP) prices were higher for Dictionary B than for Dictionary A ($t = 7.11$, $p < .001$), but in separate evaluation WTP values were higher for Dictionary A than for Dictionary B ($t = 1.69$, $p = .1$). The PR was highly significant ($t = 4.56$, $p < .001$).[2] Note that this PR occurred between

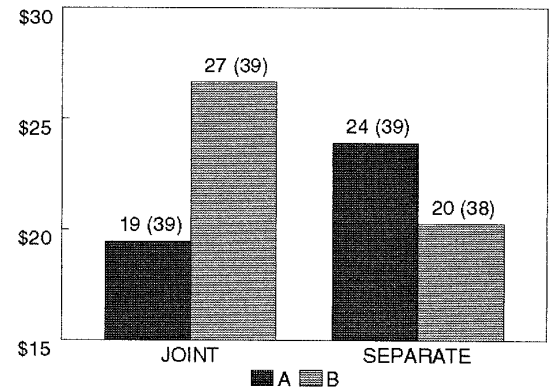

FIG. 1. Mean WTP values for Dictionary A and Dictionary B in Study 1. The numbers in parentheses indicate numbers of participants.

conditions that shared a constant, WTP, scale; the only difference lay in whether the stimulus options were evaluated jointly or separately.

Joint-separate evaluation PRs have been documented in other contexts as well. One of the original demonstrations of joint-separate evaluation PRs was provided by Bazerman, Loewenstein, and White (1992). Participants read a description of a dispute between themselves and their neighbor and then evaluated different potential resolutions of the dispute. Among the various resolution options were the following two:

A: $600 for self and $800 for neighbor

B: $500 for self and $500 for neighbor

In joint evaluation, participants were presented with pairs of options, such as the one listed above, and asked to indicate which was more acceptable or more satisfying. In separate evaluation, participants were presented with these options one at a time, and asked to indicate on a rating scale how acceptable or how satisfying each option was. Overall, rates of preference reversal between joint and separate evaluations were quite high. For example, of the two options listed above, most participants rated Option A more favorably in joint evaluation, but most rated Option B more favorably in separate evaluation. Bazerman, Schroth, Pradhan, Diekmann, and Tenbrunsel (1994) replicated this PR with business students in the context of hypotheti-

[1] To prevent their undue influences, extreme WTP values, defined here as those at least three standard deviations from the mean, were excluded prior to analysis. This footnote applies to all the studies reported in this article.

[2] To assess the significance of a joint-separate evaluation PR, one needs to compare the difference between the valuations of A and B in joint evaluation with that in separate evaluation. Note that the difference in joint evaluation is within subjects and that in separate evaluation is between subjects. To meet this need, the following t statistic is used: $t = ((M_{JA} - M_{JB}) - (M_{SA} - M_{SB}))/[(S_J^2/N_J + S_{SA}^2/N_{SA} + S_{SB}^2/N_{SB})]^{\frac{1}{2}}$, where M_{JA}, M_{JB}, M_{SA} and M_{SB} are means for A and for B in joint evaluation and means for A and for B in separate evaluation, respectively; S_J^2, S_{SA}^2 and S_{SB}^2 are variances; N_J, N_{SA}, and N_{SB} are numbers of participants in the joint, and the two separate-evaluation conditions, respectively. I thank Jimmy Ye for his help on this statistic.

cal job offers which differed in terms of (a) salary for oneself and (b) salary for others, and differed in terms of (a) salary for oneself and (b) fairness of the grievance procedure of the company (see Bazerman *et al.*, 1994, for details). Similar PRs have also been obtained by Hsee (1993) in the context of salary preferences and by Lowenthal (1993) in the context of political candidates preferences.

It should be noted that joint-separate evaluation PRs are different from the observation that effects revealed in a within-subject design may disappear in a between-subject design (e.g., Fox & Tversky, 1995). In a joint-separate evaluation PR, the preference revealed in joint evaluation does not disappear in separate evaluation; it *reverses* itself.

Joint-separate evaluation PRs cannot be easily accounted for by theories designed to explain choice-pricing and choice-matching PRs. The standard explanation for choice-pricing PRs is the compatibility principle (Slovic, Griffin, & Tversky, 1990). According to this principle, a given attribute will carry more weight in a response that is on the same scale as this attribute than in a response that is on a different scale. For example, monetary attributes will loom larger if the evaluation is made on a monetary scale, such as in pricing, than if it is made in terms of choice. Evidently, this principle is concerned with PRs involving different evaluation scales and is not applicable to joint-separate evaluation PRs. The standard explanation for choice-matching PRs is the prominence principle (Tversky *et al.*, 1988; see also Fischer & Hawkins, 1993). It posits that the most prominent attribute of the stimulus options has a greater weight in choice than in matching. However, there are substantial differences between choice and joint evaluation, and between matching and independent evaluation. For example, in matching the evaluator is exposed to both stimulus options and performs careful trade-off analyses (Tversky *et al.*, 1988); in separate evaluation the evaluator is presented with only one option and cannot perform trade-off analyses. Moreover, as will be demonstrated later in this article, joint-separate evaluation PRs can be turned "on" and "off" by varying the relative evaluability of the attributes, even if the relative prominence of those attributes remains the same.

THE EVALUABILITY HYPOTHESIS

In this section I propose an explanation for joint-separate evaluation PRs, called the evaluability hypothesis.[3] Unless otherwise specified, the discussion below assumes that there are two options to be evaluated and that the two options vary on two attributes.

	Attribute 1	Attribute 2
Option A:	a_1	a_2
Option B:	b_1	b_2

Also assume that Option A is superior to Option B on one of the attributes and Option B superior to Option A on the other attribute. The two options in Study 1 comply with this pattern. The differences between the two dictionaries can be interpreted as follows:

	Entries	Defects
Dictionary A:	10,000	no
Dictionary B:	20,000	yes

Dictionary A was superior on the Defects attribute and Dictionary B superior on the Entries attribute.

According to the evaluability hypothesis, joint-separate evaluation PRs occur because one of the attributes involved in the stimulus options is *hard to evaluate independently* and the other attribute is relatively *easy to evaluate independently.* To say that an attribute is hard to evaluate independently means that the evaluator does not know how good a given value on the attribute is without comparisons; to say that an attribute is easy to evaluate independently means that the evaluator knows how good the value is. In study 1, for example, the Entries attribute was hard to evaluate independently. Without something to compare with, most students would not know how good a dictionary with 10,000 entries (or with 20,000 entries) is. On the other hand, the Defects attribute was relatively easy to evaluate independently. Even without a direct comparison, most people would find a defective dictionary unattractive, and a like-new dictionary attractive.

The relative impact between the hard-to-evaluate and the easy-to-evaluate attributes will vary depending on the mode of the evaluation. In separate evaluation, because people do not know how to evaluate an option's value on the hard-to-evaluate attribute, they have to base their evaluation chiefly on the easy-to-evaluate attribute alone. For example, in Study 1, because those evaluating only one of the dictionaries would not know how good its number of entries was,

[3] Loewenstein, Blount, and Bazerman (1994) proposed a similar account of the joint-separate evaluation preference reversal which they cast in terms of attribute ambiguity rather than (in)evaluability. Although we developed our ideas independently, I have benefited from discussions with those authors.

they would be forced to base their evaluation of the dictionary on its cosmetic condition alone. In joint evaluation, people could compare one option against the other, and this comparison would increase the evaluability of the otherwise hard-to-evaluate attribute. For example, in the joint evaluation condition of Study 1, respondents could compare one dictionary against the other, and through this comparison they would recognize that a dictionary with 20,000 entries was relatively good and one with only 10,000 entries not as good. In short, separate evaluation is determined primarily by the easy-to-evaluate attribute and not by the hard-to-evaluate attribute, whereas joint evaluation is influenced by both the hard-to-evaluate and the easy-to-evaluate attributes.

Based on the preceding discussion, the evaluability hypothesis can be stated as follows: *When two stimulus options involve a trade-off between a hard-to-evaluate attribute and an easy-to-evaluate attribute, the hard-to-evaluate attribute has a lesser impact in separate evaluation than in joint evaluation, and the easy-to-evaluate attribute has a greater impact.* In terms of Study 1, this hypothesis implies that the Entries attribute had a lesser impact in separate evaluation than in joint evaluation, and the Defects attribute had a greater impact.

The evaluability hypothesis makes a specific prediction for the direction of joint-separate evaluation PRs. Because the hard-to-evaluate attribute loses impact and the easy-to-evaluate attribute gains impact from joint evaluation to separate evaluation, the direction of any joint-separate evaluation PRs will always be *from the option superior on the hard-to-evaluate attribute in joint evaluation to the option superior on the easy-to-evaluate attribute in separate evaluation.* Indeed, the PR observed in Study 1 conforms to this pattern. (Of course, in order for a PR to happen, the option superior on the hard-to-evaluate attribute must be preferred in joint evaluation; otherwise there would be no room for a PR. Unless otherwise specified, the above condition is assumed to be true in the rest of this article.)

The evaluability hypothesis is also consistent with the finding of the self-neighbor study by Bazerman *et al.,* (1992). In that study, the two outcomes mentioned above can be interpreted as varying on two attributes: (a) payoff to oneself and (b) whether this payoff equaled the payoff to one's neighbor:

	Payoff	Equality
Option A:	$600	unequal
Option B:	$500	equal

The payoff attribute was presumably relatively hard to evaluate independently. In contrast, the equality attribute was relatively easy to evaluate independently; even without a direct comparison, most people would find an equal settlement appealing and an unequal settlement unappealing. Again, the PR observed in that study was in the direction predicted by the evaluability hypothesis, that is, from Option A (superior on the hard-to- evaluate attribute) in joint evaluation to Option B (superior on the easy-to-evaluate attribute) in separate evaluation. Similar analyses can be applied to other joint-separate evaluation PR findings (e.g., Bazerman, *et al.*, 1994; Hsee, 1993; Lowenthal, 1993).

So far, the evaluability hypothesis has only been used to make post hoc explanations for already observed PRs. The following studies were designed to test whether the evaluability hypothesis is capable of making predictions. These studies each involved options that varied on a hard-to-evaluate attribute and an easy-to-evaluate attribute. The evaluability hypothesis was used to predict the direction of a PR. Study 2 used naturally occurring hard-to-evaluate and naturally occurring easy-to-evaluate attributes. In Studies 3 and 4, whether an attribute was hard or easy to evaluate was manipulated empirically.

STUDY 2

Study 2 differed from Study 1 in two major respects. First, as mentioned earlier, Study 2 was designed to test the evaluability hypothesis rather than simply to demonstrate a PR. Second, in Study 1 as well as in the other studies reviewed previously, the evaluability of an attribute was confounded with the continuous/dichotomous nature of the attribute; the hard-to-evaluate attribute was always a continuous variable and the easy-to-evaluate attribute always a dichotomous variable. In Study 2, both the hard-to-evaluate and the easy-to-evaluate attributes were continuous variables.

Method

Design and stimuli. This study involved the evaluations of two hypothetical job candidates for a computer programmer position. The programmer was expected to use a computer language called KY. The two candidates were:

	Candidate A	Candidate B
Education:	B.S. in computer science from UIC	B.S. in computer science from UIC
GPA from UIC:	4.9	3.0
Experience with KY:	has written 10 KY programs in the last 2 years	has written 70 KY programs in the last 2 years

("UIC" stands for the University of Illinois at Chicago. The participants were students of that university and knew the abbreviation. GPA at UIC is on a 5-point scale.)

Note that the two candidates differed on two attributes—GPA and Experience. Both are continuous variables. For ease of discussion later, let us summarize the differences between the candidates in the following format:

	Experience	GPA
Candidate A:	10 KY programs	4.9
Candidate B:	70 KY programs	3.0

As in Study 1, the questionnaire for this study had three between-subject versions, joint-evaluation, separate-evaluation-A, and separate-evaluation-B. In all three versions, participants were asked to imagine that they were the owner of a consulting firm, that they were looking for a computer programmer to use a computer language called KY, and that they planned to pay the person between $20,000 and $40,000 per year. In the joint-evaluation condition, participants evaluated both candidates. In each separate-evaluation condition, they evaluated only one of the candidates. The evaluation scale was constant across the three versions—willingness-to-pay salary.

Measure of evaluability. To assess which attribute was hard to evaluate independently and which was easy, participants in the two separate-evaluation conditions were asked the following questions after they had indicated their WTP salaries for the candidate: (a) "Do you have any idea how good a GPA of 4.9 (3.0) from UIC is?" and (b) "If someone has written 10 (70) KY programs in the last 2 years, do you have any idea how experienced he/she is with KY?" (The numbers preceding the parentheses were for the separate-evaluation-A condition and those in the parentheses were for the separate-evaluation-B condition.) To answer each question, participants would choose among four options, ranging from (1) = "I don't have any idea." to (4) = "I have a clear idea." These options served as an evaluability scale, where a greater number indicated greater evaluability.

Participants and procedure. Respondents were 112 college students from the University of Illinois at Chicago. They randomly received one of the three versions of the questionnaire and completed it individually. Upon completion each participant received a candy bar as compensation.

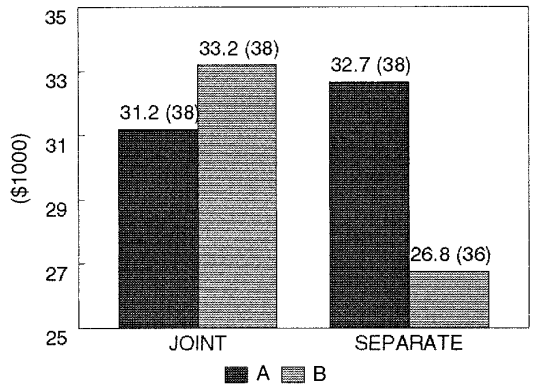

FIG. 2. Mean WTP salaries for Candidate A and Candidate B in Study 2. The numbers in parentheses indicate numbers of participants.

Results and Discussion

Measure of evaluability. The mean evaluability score for GPA was 3.7 and that for Experience was 2.1. The difference was significant ($t = 11.79$, $p < .001$). These results established that GPA was a relatively easy-to-evaluate attribute and Experience a relatively hard-to-evaluate attribute.

Willingness-to-pay values. According to the evaluability hypothesis, there was likely to be a joint-separate evaluation PR, because one of the attributes involved in the stimulus options (Experience) was hard to evaluate independently and the other attribute (GPA) relatively easy. Given that Candidate A was superior on GPA and Candidate B superior on Experience, the direction of the PR would be from Candidate B in joint evaluation to Candidate A in separate evaluation.

The results, summarized in Fig. 2 were consistent with these predictions. There was a significant PR between joint and separate evaluations ($t = 4.94$, $p < .001$). In joint evaluation, WTP salaries were higher for Candidate B than for Candidate A ($t = 1.65$, $p = .1$). In separate evaluation WTP values were higher for Candidate A than for Candidate B ($t = 5.50$, $p < .001$).

This study yields two important implications. First, joint-separate evaluation PRs exist not only when one attribute is dichotomous and the other attribute continuous, but also when both attributes are continuous. Second, this study shows that the evaluability hypothesis is not only able to provide post-dictions for already-observed PRs, but also able to provide predictions.

STUDY 3

In all of the studies discussed thus far, whether an attribute was hard or easy to evaluate independently was

a characteristic of the attribute per se and was never manipulated empirically. In the two studies described below, the evaluability of an attribute was manipulated empirically. As mentioned earlier, the evaluability hypothesis asserts that joint-separate evaluation PRs occur because one of the attributes involved in the stimulus options is hard to evaluate independently while the other attribute relatively easy, and the relative impact of the two attributes changes from joint evaluation to separate evaluation. It implies that if both attributes are hard to evaluate independently, or if both easy to evaluate independently, then the relative impact of the two attributes will not change between the two evaluation modes, and there will be no PR.

If the foregoing analysis is correct, then a PR can be turned either "on" or "off" by varying the relative evaluability of the attributes. Study 3 and Study 4 were designed to test this intuition.

Study 3 involved two evaluability conditions, Hard/Hard and Hard/Easy: In the Hard/Hard condition, both attributes were hard to evaluate independently; in the Hard/Easy condition, one was hard and one easy. The evaluability of an attribute was manipulated depending on whether or not participants were informed of the meanings of the attribute. In the Hard/Hard condition, the values on both attributes (Clarity and Warranty of a TV) were meaningless numbers. In the Hard/Easy condition, participants were told that the Warranty rating meant the length of the warranty; presumably the Warranty attribute would be easier to evaluate independently once participants knew its meanings. The prediction for this study was that there would no PR between joint and separate evaluations in the Hard/Hard condition and that there would be one in the Hard/Easy condition.

Method

Design and stimuli. Study 3 involved the evaluations of two hypothetical TVs; they varied on two attributes, Clarity and Warranty:

	Clarity	Warranty
TV A:	90	9
TV B:	40	18

The questionnaire for this study had three versions and each included two parts. They constituted 3 Evaluation Mode × 2 Evaluability conditions. In all versions, participants were asked to assume that they were shopping for a basic 20″ color TV, and that most such TVs would cost around $200. Participants were also asked to assume that they were in a store where the salespeople knew nothing about TVs, and that the tag on the TV(s) contained two indices, Clarity and Warranty. In the joint-evaluation condition, participants indicated their WTP prices for both TVs and in each separate-evaluation condition for only one of the TVs.[4]

The two Evaluability conditions were Hard/Hard and Hard/Easy. In the Hard/Hard condition, both the Clarity and the Warranty ratings were meaningless numbers and hence both hard to evaluate independently. Participants were simply told that Clarity reflected how clear the picture was, that Warranty reflected how good the warranty was, and that for both indices, the higher the number, the better. In the Hard/Easy condition, Clarify remained hard to evaluate, but Warranty was made relatively easy to evaluate by telling participants that the Warranty rating indicated the length, in months, of the warranty.

The two Evaluability conditions were presented within-subjects. Because the Hard/Easy condition contained information not available in the Hard/Hard condition but not vice versa, the Hard/Hard condition always preceded the Hard/Easy condition.

Participants and procedure. Respondents were 98 college students from the University of Chicago who completed multiple questionnaires and received a cash payment. Each participant received one of the three versions of the questionnaire and completed it individually.

Results and Discussion

According to the evaluability hypothesis, there would be no PR in the Hard/Hard condition, and there was likely to be a PR in the Hard/Easy condition. Because TV A was superior on the hard-to-evaluate attribute (Clarity) and TV B superior on the easy-to-evaluate attribute (Warranty), the direction of the PR would be from TV A in joint evaluation to TV B in separate evaluation.

The results, summarized in Fig. 3, confirmed these predictions. In the Hard/Hard condition, there was no PR: WTP values were higher for TV A than for TV B in both joint evaluation ($t = 5.5$, $p < .001$) and separate evaluation (although the difference in separate evaluation was not significant). In the Hard/Easy condition, there was a significant PR ($t = 3.47$, $p < .01$): In joint evaluation, WTP values were higher for TV A than for TV B ($t = 4.33$, $p < .001$), but in separate evaluation

[4] In Study 3 and Study 4, participants in the joint-evaluation condition first indicated whether their WTP price was higher for A or for B before indicating a specific WTP price for each option. Two participants in Study 3 and four in Study 4 gave contradictory responses, i.e., said that they were willing to pay more for one option but gave a higher WTP price for the other. These responses were excluded.

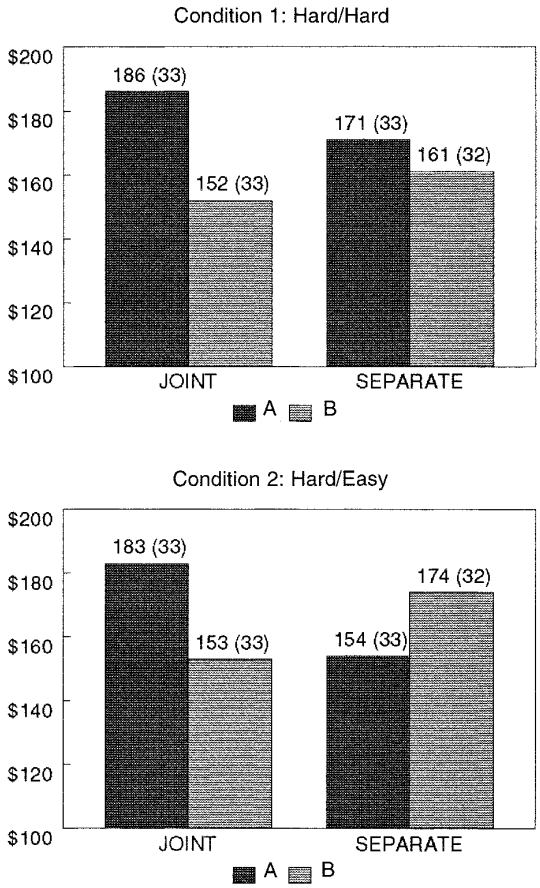

FIG. 3. Mean WTP values for TV A and TV B in Study 3. The numbers in parentheses indicate numbers of participants.

WTP values were higher for TV B than for TV A (t = 1.56, p = .12).

STUDY 4

Study 4 was a replication of Study 3 with the following main differences. First, like Study 3, Study 4 had two Evaluability conditions. Instead of Hard/Hard and Hard/Easy, the two conditions were Hard/Easy and Easy/Easy. It was predicted that there would be a PR in the Hard/Easy condition, but no PR in the Easy/Easy condition. Second, the two Evaluability conditions were between-subjects rather than within-subjects. Third, evaluability was manipulated differently in Study 4 than in Study 3. In the Hard/Easy condition, the hard-to-evaluate attribute was an unfamiliar variable (total harmonic distortion of a CD changer); in the Easy/Easy condition, the possible range of the total-harmonic-distortion attribute was provided. It is expected that an unfamiliar attribute will become relatively easier to evaluate if one knows the range of the attribute and hence knows where the focal value falls in this range. Finally, a manipulation check was used in Study 4 to verify the effectiveness of the evaluability manipulation.

Method

Design and stimuli. This study involved the evaluations of two CD changers (i.e., multiple compact disc players):

	CD Changer A	CD Changer B
Brand:	JVC	JVC
CD capacity:	can hold 5 CDs	can hold 20 CDs
Sound quality:	THD = .003%	THD = .01%
Warranty:	1 year	1 year

Note that the two CD changers varied on two attributes: CD-capacity and sound quality; the latter was indexed by THD. It was explained to participants in all conditions that THD stands for total harmonic distortion, and that the smaller the THD, the better the sound quality. For ease of discussion, let us summarize the differences between the two CD changers as follows:

	THD	CD Capacity
CD Changer A:	.003%	5 CDs
CD Changer B:	.01%	20 CDs

The questionnaire for this study had six between-subject versions; they constituted 3 Evaluation Mode × 2 Evaluability conditions. In all conditions, participants were asked to assume that they were shopping for a CD changer in a department store and that the price of a CD changer would range from $150 to $300. In the joint-evaluation condition participants indicated their WTP prices for both CD changers; in each separate-evaluation condition, for only one of the CD changers.

The two Evaluability conditions were Hard/Easy and Easy/Easy. In the Hard/Easy condition, participants received no other information about either THD or CD-Capacity than described previously. It was assumed that THD was hard to evaluate independently and CD-capacity relatively easy. Without a comparison, most students would not know whether a given THD rating (e.g., .01%) was good or bad, but they would have some idea of how many CDs a CD changer could hold and whether a CD changer that can hold 5 CDs (or 20 CDs) was good or not.

In the Easy/Easy condition, participants were given the following additional information about THD: "For most CD changers on the market, THD ratings range from .002% (best) to .012% (worst)." This information was designed to make THD easier to evaluate independently. With this information, participants in the separate-evaluation conditions would have some idea where the given THD rating fell in the range and hence whether the rating was good or bad. Participants received no additional information about CD-capacity.

Measure of evaluability. To check that the evaluability manipulation was effective, participants in the two separate-evaluation conditions were asked the following questions after they had indicated their WTP prices: "Do you have any idea how good a THD rating of .003% (.01%) is?" and "Do you have any idea how large a CD capacity of 5 (20) CDs is?" (The numbers preceding the parentheses were for the separate-evaluation-A condition and those in the parentheses were for the separate-evaluation-B condition.) As in Study 2, answers to those questions ranged from 1 to 4, greater numbers indicating greater evaluability.

Participants and procedure. Respondents were 202 college students from the University of Illinois at Chicago. They randomly received one of the six versions of the questionnaire and completed it individually. Each participant received a candy bar as compensation.

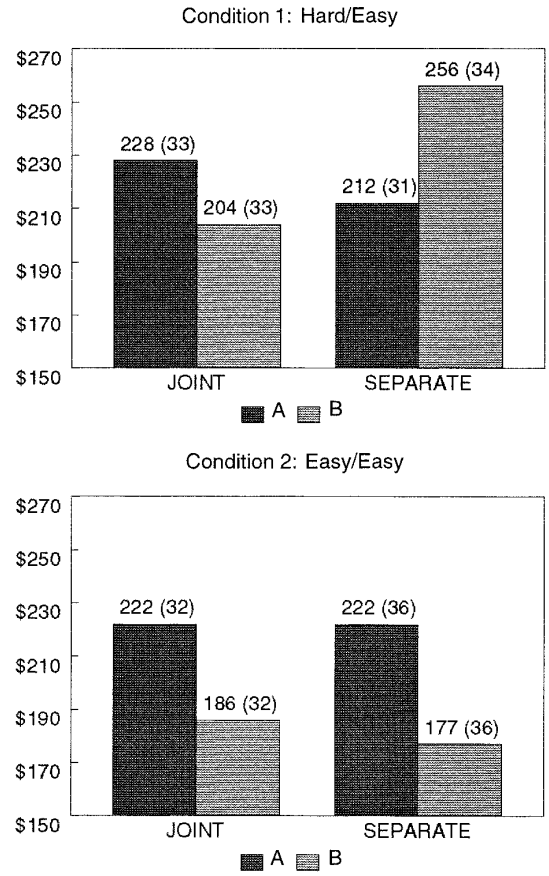

FIG. 4. Mean WTP values for CD Changer A and CD Changer B in Study 4. The numbers in parentheses indicate numbers of participants.

Results and Discussion

Measure of evaluability. Mean evaluability scores for THD and CD-capacity in the Hard/Easy condition were 1.98 and 3.25, respectively, and in the Easy/Easy condition were 2.53 and 3.22. A 2 Attribute (THD versus CD-capacity) × 2 Evaluability (Hard/Easy versus Easy/Easy) analysis of variance revealed a significant interaction effect ($F(1,135) = 9.40$, $p < .01$) and a significant main effect for Attribute ($F(1,135) = 111.79$, $p < .001$). Planned comparisons indicated that evaluability scores for THD were significantly higher in the Easy/Hard condition than in the Hard/Hard condition ($t = 2.92$, $p < .01$), suggesting that the evaluability manipulation for THD was effective. There were virtually no differences in the evaluability of CD-capacity between the Easy/Hard and the Easy/Easy conditions.

Willingness-to-pay values. Based on the evaluability hypothesis, the following predictions were made: A PR was likely to occur in the Hard/Easy condition, but not in the Easy/Easy condition. In the Hard/Easy condition, because CD Changer A was superior on the hard-to-evaluate attribute (THD) and CD Changer B superior on the easy-to-evaluate attribute (CD-capacity), the direction of the PR in that condition would be from CD Changer A in joint evaluation to CD Changer B in separate evaluation.

The results, summarized in Fig. 4, confirmed these predictions. In the Hard/Easy condition, there was a significant PR ($t = 3.32$, $p < .01$), and the direction of the PR was consistent with the evaluability hypothesis: In joint evaluation WTP values were higher for CD Changer A than for CD Changer B ($t = 1.96$, $p = .06$), but in separate evaluation WTP values were higher for CD Changer B than for CD Changer A ($t = 2.70$, $p < .01$). In the Easy/Easy condition, the PR disappeared ($t < 1$, n.s.). WTP values were higher for CD Changer A than for CD Changer B in both joint evaluation ($t = 2.81$, $p < .01$) and separate evaluation ($t = 2.92$, $p < .01$).

Study 4 corroborates Study 3 by showing that the

presence of a joint-separate evaluation PR depends on the relative evaluability of the attributes involved in the stimulus options. A PR is likely to occur if one attribute is hard to evaluate independently and the other relatively easy to evaluate independently. The chance of a PR is greatly mitigated if both attributes are hard to evaluate independently (such as in the Hard/Hard condition of Study 3), or if both attributes are easy to evaluate independently (such as in the Easy/Easy condition of Study 4).

GENERAL DISCUSSION

Preference reversals as traditionally studied are usually between conditions that involve different evaluation scales, e.g., acceptability versus pricing. The present paper concerns itself with preference reversals between conditions that share the same evaluation scale, but differ in the way the options are evaluated—either jointly or separately. According to the evaluability hypothesis, joint-separate evaluation PRs occur because one of the attributes involved in the options is hard to evaluate independently while the other attribute is relatively easy. Study 1 demonstrated a joint-separate evaluation PR when the easy-to-evaluate attribute was a dichotomous variable and the hard-to-evaluate attribute a continuous variable. Study 2 showed that a PR could occur even if both attributes were continuous. In Study 3 and Study 4, the evaluability of an attribute was manipulated empirically. A PR emerged when one attribute was hard to evaluate and the other easy to evaluate, and it disappeared if both attributes were hard to evaluate or both easy to evaluate. The findings of these studies provide consistent support for the evaluability hypothesis.

Several potential questions about this research need to be addressed. First, what determines whether an attribute is hard or easy to evaluate independently? My speculation is that it depends on how much knowledge the evaluator has about that attribute, especially about the value distribution of that attribute. An attribute will be hard to evaluate independently if the evaluator does not know its distribution information, such as the possible values of the attribute, its best and worst values, and so forth. Without such knowledge, the evaluator will not know where a given value on that attribute lies in relation to the other values on the attribute and hence will not know how to evaluate it. Indeed, the evaluability manipulations in Study 3 and Study 4 were based on these intuitions. For example, in the Hard/Easy condition of Study 4, THD was hard to evaluate independently because most participants did not have any distribution information of THD. In the Easy/Easy condition of that study, THD became relatively easier to evaluate independently because participants were given some distribution information of that attribute, namely, its range. With this information, people would not know where the THD value of the given option fell in the range and hence whether it was good or not.

Whether or not an attribute is hard to evaluate independently may be related to the certainty with which one can evaluate its values. Mellers, Richards, and Birnbaum (1992) found that an adjective carries a lesser impact if the evaluation of the adjective is uncertain (and has a large variance) than if it is certain (and has a small variance). It is possible that for hard-to-evaluate attributes there is greater uncertainty in judging their values in separate evaluation than in joint evaluation, and therefore these attributes have a lesser impact on separate evaluation than on joint evaluation. For easy-to-evaluate attributes, the uncertainty is likely to be low regardless of the evaluation mode, and hence these attributes will have consistent impact on separate and joint evaluations.

Another question about this research concerns its relationship with the prominence principle (Tversky *et al.*, 1988). Although that principle was originally formulated to explain choice-matching PRs, one may interpret it more generally to mean that the most prominent attribute of an option will carry more weight in joint evaluation than in separate evaluation. This interpretation of the prominence principle is consistent with the results of Studies 1 and 2, but not with those of Study 3 or Study 4. If a joint-separate evaluation PR occurs simply because one attribute is more prominent than the other, then the Evaluability manipulations in Study 3 and Study 4 should not have affected the existence of a PR, because these manipulations would not alter the relative prominence of the attributes.

Finally, one may wonder if there are other explanations for joint-separate evaluation PRs than the evaluability hypothesis. The answer is probably yes. Generally speaking, the evaluability hypothesis is most applicable to PRs where the options vary on two distinct attributes, and one of the attributes is markedly harder to evaluate independently than the other attribute. PRs of other forms may or may not be explained by the evaluability hypothesis, and there are also other possible explanations. For example, when two options involving a tradeoff between two attributes, and the values of the options are of the same sign on one attribute but of different signs on the other attribute, a PR may emerge between joint and separate evaluations (see Hsee, 1994, for a demonstration). This type of PR

can be explained in terms of the different evaluation models people use in joint versus in separate evaluation and the curvilinearity of the utility functions of the attributes (Bazerman *et al.*, 1992; Hsee, 1994). Generally speaking, people use an additive difference model (Tversky, 1969) in joint evaluation and an additive model in separate evaluation, and most attributes have a concave utility function in the positive domain and a convex utility function in the negative domain (e.g., Kahneman & Tversky, 1979). From these assumptions it can be derived mathematically that the attribute on which the stimulus options are of the same sign will have a lesser impact in separate evaluation than in joint evaluation (Bazerman *et al.*, 1992; Hsee, 1994). See Bazerman *et al.* (1992) for an alternative explanation of the result of their self–neighbor study from this perspective.

As another example, consider a study reported by Irwin, Slovic, Lichtenstein and McClelland (1993). Participants stated their WTP values for various options, including (a) improvement in air quality in Denver and (b) improvement in a VCR. In joint evaluation WTP values were higher for improvements in air quality, but in separate evaluation WTP values were higher for improvements in consumer products (see also Kahneman & Ritov, 1994, for similar findings). This type of PR was quite different from the one explored in the present research; for example, the options in Irwin *et al.* (1994) were of two entirely different categories (air quality versus consumer products) and they shared no common attributes. Irwin *et al.* (1993) considered their finding as an instance of choice-judgment PR and explained it using the prominence principle. Another *possible* explanation is related to elastic justification (Hsee, 1995, 1996), the notion that one's decision will be influenced more by unjustifiable considerations when there is ambiguity as to what one should do than when there is not. It is reasonable to assume that improving one's VCR is a less justifiable action than improving air quality in Denver, and that separate evaluation entails more ambiguity as to what one should do than joint evaluation. Thus, the results of Irwin *et al.* (1993) may be an incidence of elastic justification.

I shall conclude this article with a discussion of some prescriptive implications. In real life, people often face the option of engaging either in joint evaluation or in separate evaluation. For example, when a manufacturer launches a new product, he may face the choice of either running a comparative advertisement against a rival product or advertising his product by itself. In the former scenario the focal and the rival products will likely be evaluated jointly, in the latter scenario, separately. Similarly, when a person is applying for a job, she may have the option of either being interviewed at the same time as another job candidate or being interviewed on a different day than the other candidate. In the same-time scenario, the candidate is likely to be evaluated jointly with the other candidate. In the different-day case, especially if the two interview days are scheduled far enough apart, she is likely to be evaluated separately from her competitor. Which option should the manufacturer adopt? Which should the job candidate adopt? According to the findings of this research, the answers to these questions depend on the type of attributes on which one excels over one's rival. If one is superior to one's rival on hard-to-evaluate attributes and inferior on easy-to-evaluate attributes, one should try to create a joint evaluation environment so as to facilitate direct comparison. If the reverse is true, one should try to be evaluated separately from one's rival.

The present research shows that when two options involving a trade-off between a hard-to-evaluate attribute and an easy-to-evaluate attribute are evaluated, preference between these options may change depending on whether these options are presented jointly or separately. More important, the direction of this change can be predicted, and can even be manipulated.

REFERENCES

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly,* 37, 220–240.

Bazerman, M. H., Schroth, H. A., Shah, P. P., Diekmann, K. A., & Tenbrunsel, A. E. (1994). The inconsistent role of comparison others and procedural justice in reactions to hypothetical job descriptions: Implications for job acceptance decisions. *Organizational Behavior and Human Decision Processes,* 60, 326–352.

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

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics,* 585–603.

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review,* 94, 236–254.

Grether, D. M., & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *American Economic Review,* 69, 623–638.

Hsee, C. K. (1993). *When trend of monetary outcomes matters: Separate versus joint evaluation, and judgment of feelings versus choice.* Unpublished manuscript, University of Chicago.

Hsee, C. K. (1995). Elastic justification: How tempting but task-irrelevant factors influence decisions. *Organizational Behavior and Human Decision Process,* 62, 330–337.

Hsee, C. K. (1996). Elastic justification: How unjustifiable factors

influence judgments. *Organizational Behavior and Human Decision Processes,* 66, 122–129.

Hsee, C. K. (1994). *Preference reversal between joint and separate evaluations.* Working paper, The University of Chicago.

Irwin, J. R., Slovic, P., Lichtenstein, S., & McClelland, G. H. (1993). Preference reversals and the measurement of environmental values. *Journal of Risk And Uncertainty,* 6, 5–18.

Kahneman, D., & Ritov, I. (1994). Determinants of stated willingness to pay for public goods. A study in the headline method. *Journal of Risk And Uncertainty,* 9, 5–38.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* 47, 263–291.

Lichtenstein, S., & Slovic, P. (1971). Reversal of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology,* 89, 46–55.

Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1994). *Reversals of Preference Between Independent and Simultaneous Evaluation of Alternatives.* Working paper, Carnegie Mellon University.

Lowenthal, D. (1993). *Preference reversals in candidate evaluation.* Working paper, Carnegie Mellon University.

Mellers, B. A., Richards, V., & Birnbaum, M. H. (1992). Distributional theory of impression formation. *Organizational Behavior and Human Decision Process,* 51, 313–343.

Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in judgment and choice. In R. M. Hogarth (Ed.), *Insights in Decision Making: Theory And Applications.* Chicago: Univ. of Chicago.

Slovic, P., & Lichtenstein, S. (1968). The relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology Monographs,* 78, (3, Pt.2).

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review,* 76, 31–48.

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

Received: January 9, 1996