31. Attribute Evaluability

Its Implications for Joint–Separate Evaluation Reversals and Beyond[1]

Christopher K. Hsee

ABSTRACT. The evaluability hypothesis posits that when two objects are evaluated separately, whether a given attribute of the objects can differentiate the evaluations of these objects depends on whether the attribute is easy or difficult to evaluate independently. The article discusses how the evaluability hypothesis explains joint–separate evaluation reversal, which is the phenomenon that the rank order of the evaluations of multiple objects changes depending on whether these objects are evaluated jointly or separately. The article presents empirical evidence for the evaluability hypothesis. The final section of the article discusses implications of the hypothesis for issues beyond reversals – in particular for inconsistencies between decisions and their consequences. Decisions are typically made in the joint evaluation mode, and the outcome of a decision is usually experienced (or "consumed") in the separate evaluation mode. Thus, reversals between joint and separate evaluation may manifest themselves in decision–consumption inconsistencies.

I. INTRODUCTION

All judgments and decisions are made in one (or some combination) of two basic modes: joint and separate. In the joint evaluation (JE) mode, people are exposed to multiple objects simultaneously and evaluate these objects comparatively. In the separate or single evaluation (SE) mode, people are exposed to only one object and evaluate it in isolation. For example, when shopping for a piano at a music instrument store, we are usually in the joint evaluation (JE) mode because there are typically many pianos for us to compare. On the other hand, when we debate whether to bid for a piano at an estate auction where there are no other pianos, we probably think about this particular piano alone and are therefore in the separate or single evaluation (SE) mode. Of course, our evaluations sometimes fall somewhere between the two modes. For example,

[1] This chapter incorporates new material from Hsee (1998) and Hsee et al. (1999). I thank the following people (in alphabetical order of their last names) for their comments: Scott Jeffrey, Josh Klayman, Howard Kunreuther, George Loewenstein, Cade Massey, Joe Nunes, Eldar Shafir, Paul Slovic, Jack Soll, John Wright, and George Wu.

This article is a revised version of one that appeared in *Organizational Behavior and Human Decision Processes*, 67:3, 247–57.

when evaluating job candidates, we are partly in the JE mode and partly in the SE mode – in the JE mode to the extent that we compare one candidate with another candidate and in the SE mode to the extent that we focus on one candidate at a time. Strictly speaking, the evaluation mode is a continuum, with JE on one end and SE on the other. For the sake of simplicity, this article concerns itself mainly with the two ends of the continuum.[2]

Will a set of objects (say, two objects) be evaluated differently between JE and SE? The answer is yes. Sometimes, even the rank order of the evaluations of the two objects reverses itself between the two evaluation modes, that is, one object is favored in JE and the other object is favored in SE. This phenomenon will be referred to as joint–separate evaluation reversal or simply JE–SE reversal.

The present article is organized as follows: The next section presents evidence showing JE–SE reversals. The section that follows discusses the evaluability hypothesis, an explanation for the reversal. The succeeding section provides empirical support for that hypothesis. The last several sections of the article examine the implications of the evaluability hypothesis for issues beyond JE–SE reversals.

II. EVIDENCE FOR JOINT–SEPARATE EVALUATION REVERSALS

1. Dictionary Study

This study illustrates the JE–SE reversal effect. The study involved the evaluation of two hypothetical second-hand music dictionaries as follows:

	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.

Respondents (116 students from the University of Chicago and the University of Illinois at Chicago) were assigned to one of three conditions – JE, SE–A and SE–B.[3] In each condition, participants were asked to assume that they were music majors looking for a music dictionary in a used book store and planned to spend between $10 and $50. In the JE condition, participants were told that there

[2] The term *separate evaluation* or *SE* refers both to (1) situations in which different objects are presented to and evaluated by different individuals so that each individual sees and evaluates only one object and to (2) situations in which different objects are presented to and evaluated by the same individuals at different times so that each individual evaluates only one object *at a given time*. The former situations are pure separate evaluation conditions. The latter situations involve some joint evaluation flavor because individuals evaluating a later object may recall the previous object(s) and make a comparison. In the studies reported in this paper, the separate evaluation conditions were like the former situations.

[3] For the sake of consistency, in all the studies discussed in this article I use the letter "A" to label the option favored in SE and the label "B" to label the other option.

Table 31.1. Mean WTP Values for the Two Dictionaries in the Dictionary Study

Evaluation Mode	Dictionary A	Dictionary B
Joint	$19	$27
Separate	$24	$20

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 SE 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.

Table 31.1 summarizes the results:[4] There was a clear reversal in willingness to pay (WTP) between JE and SE. In JE, WTP values were higher for Dictionary B ($t = 7.11$, $p < .001$), but in SE, WTP values were higher for Dictionary A ($t = 1.69$, $p < 0.1$).

2. Other Evidence

Joint–separate evaluation reversals have been documented in other contexts as well. The original demonstration of JE–SE reversal was provided by Bazerman, Loewenstein and White (1992) in the context of dispute resolution. Those authors found that between an option that entailed a high payoff to oneself and looked unfair (e.g., $600 to self and $800 to the other side) and an option that entailed a lower payoff to oneself and looked fair (e.g., $500 to both sides), the former option was favored in JE and the latter option was favored in SE. Bazerman et al. (1994) replicated these results with business students in the context of hypothetical job offers. In the context of political candidate preferences, Loewenthal (1993) found that in JE people preferred a candidate who could bring 5,000 jobs to the district but had a minor conviction in the past to one who could bring only 1,000 jobs but had a clean history; in SE, the preference was reversed. Similar JE–SE reversals have been obtained by Hsee (1993) in the context of salary preferences and by Nowlis and Simonson (1994) in the context of consumer products.

Joint–separate evaluation reversals suggest that switching from one evaluation mode to the other can change the attractiveness of one object *relative* to the other object. It should be noted, however, that varying evaluation mode can also produce other types of changes. In a recent study, for example, Hsee and Leclerc (1998) showed that, under a set of predictable conditions, switching from one evaluation mode to the other increases the attractiveness of *both* objects even though their relative attractiveness stays the same. In the present

[4] To prevent their undue influences, willingness-to-pay values more than three standard deviations from the mean were excluded from analysis. This footnote applies to all the studies reported in this article.

article, however, I limit my discussions to changes in *relative* attractiveness of the objects – in particular to rank-order reversal.

3. Differences between JE–SE Reversal and Choice–Judgment Preference Reversals

Joint–separate evaluation reversals are not the same as the traditionally studied choice–judgment preference reversals. Of the many forms of choice-judgment reversals documented in the decision literature, the most widely studied are probably the choice-pricing reversal and the choice-matching reversal. In the choice-pricing paradigm, participants either choose between two gambles or indicate their minimum selling price for the gambles (e.g., Lichtenstein and Slovic 1971, Grether and Plott 1979). Typically, participants favor low-payoff and high-probability gambles over high-payoff and low-probability gambles in choice but demand a higher minimum selling price for the high-payoff and low-probability gambles in pricing. In the choice-matching paradigm, participants either choose between two alternatives or fill in some missing value in one of the alternatives to make the two alternatives equally attractive (e.g., Tversky, Sattath, and Slovic 1988). Typically, in choice people favor the option superior on the most prominent attribute, but in matching their responses imply more even weighting between the attributes.

In both the choice-pricing and the choice-matching paradigms, reversals occur between tasks that involve different *evaluation scales* (e.g., Goldstein and Einhorn 1987, Bazerman et al. 1992). Evaluation scale refers to the dimension on which subjects' responses are elicited, whether it is about intention to accept (acceptability), about willingness to pay, about selling price, about feelings, or about something else. In the choice-pricing paradigm, the evaluation scale for choice is acceptability and that for pricing is selling price. In the choice-matching paradigm, the evaluation scale for choice is, again, acceptability and that for matching is probability or value estimation.

Although choice and judgment reversals sometimes involve different *evaluation modes*, difference in evaluation mode is not a necessary condition for those classic preference reversals. For example, in the choice-matching reversal, a typical subclass of choice–judgment reversal, both the choice response and the matching response are elicited in the JE mode.

Unlike classic choice–judgment reversals, the critical difference between the conditions that produce JE–SE reversals is evaluation mode – whether the target objects are presented and evaluated jointly or presented and evaluated separately. Whether the two evaluation modes involve the same or different evaluation scales is not essential. A reversal can occur between JE and SE even if the evaluation scale between the two evaluation modes is held constant. In the previously described dictionary study, for instance, the reversal occurred between JE and SE, even though the evaluation scale in both conditions was invariably willingness to pay.

Joint–separate evaluation reversals cannot be easily accounted for by standard theories for choice–judgment reversals. These theories rely on difference in

evaluation scale rather than difference in evaluation mode to explain preference reversals. For example, the standard explanation for choice-pricing reversals is the compatibility principle (Slovic, Griffin, and 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 a pricing task) than if it is made in terms of acceptability (such as in a choice task). Clearly, this principle is concerned with evaluation scales rather than with evaluation modes.

The standard explanation for choice-matching reversals is the prominence principle; it is a special case of the compatibility principle (Tversky et al. 1988; see also Fischer and Hawkins 1993). According to the prominence principle, people use a lexicographic rule when deciding which option to accept (choice) and use a trade-off analysis when judging a missing attribute value (matching). It is posited that the more prominent attribute is more compatible with a lexicographic decision task than with a value judgment task involving trade-off analyses. As a result, the more prominent attribute looms larger in choice than in matching. The prominence principle provides a compelling explanation for the standard choice-matching reversal, but it does not readily apply to the JE–SE reversal studied in the present research. As demonstrated previously, JE–SE reversals can take place even if the evaluation scale is held constant. In addition, there is no direct correspondence between matching and SE, or between choice and JE. For example, in the typical matching task, the evaluator is exposed to multiple stimulus objects and performs careful trade-off analyses (Tversky et al. 1988); in SE, on the other hand, the evaluator is presented with only one object and cannot perform trade-off analyses.

In summary, JE–SE reversals are different from traditionally studied choice–judgment preference reversals and cannot be readily explained by classic preference reversal theories.

III. THE EVALUABILITY HYPOTHESIS

In this section I discuss an explanation for JE–SE reversals called the evaluability hypothesis.[5] Unless otherwise specified, the discussion below assumes that there are two objects to be evaluated and that the two objects involve a trade-off along two attributes in the form of

	Object A	**Object B**
Attribute 1:	a_1	b_1
Attribute 2:	a_2	b_2

[5] Similar accounts of JE–SE reversals were proposed by Hsee (1993) in terms of "reference-dependency" of attribute evaluation; by Loewenstein, Blount, and Bazerman (1993) in terms of "attribute ambiguity," and by Nowlis and Simonson (1994) in terms of "context–dependency."

Assume also that decision makers care about the attributes and know which direction of the attribute is better. The two options in the dictionary study comply with these assumptions. The two dictionaries involved a trade-off along the following attributes:

	Dictionary A	**Dictionary B**
Number of entries:	10,000	20,000
Defects:	no	yes

According to the evaluability hypothesis, JE–SE reversals occur because one of the attributes involved in the stimulus objects is *hard to evaluate independently* and the other attribute is relatively *easy to evaluate independently.* Below I first discuss what makes an attribute easy or difficult to evaluate and then explain why it is relevant to JE–SE reversals.

1. Attribute Evaluability

To say that an attribute is hard to evaluate independently means that, even though people know which direction of the attribute is better, they do not know how good a given value on the attribute is when the value is presented alone.

Whether an attribute is hard or easy to evaluate depends on how much knowledge the decision maker has about that attribute – especially about its effective range, its neutral reference point, and its value distribution. Without such knowledge, the decision maker will not know where a given value of that attribute lies in relation to the other values of the attribute and hence will not know how to evaluate it.

To say that an attribute is hard to evaluate does not mean that the decision maker does not know the precise value on the attribute, but means that the decision maker does not know how to interpret the desirability of that value. For example, suppose that a person traveling in a foreign country is told that a particular hotel room there costs $56.78 per night. If the person is unfamiliar with the hotel rates there, then the price of that hotel, albeit precisely given, will be hard for him to evaluate independently. (For a more detailed analysis of what determines the evaluability of an attribute, see Hsee et al. 1999).

In the dictionary study, for example, the number of entries was hard to evaluate independently. Most respondents had little knowledge about the number of entries in a music dictionary. Without something to compare it with, they would not know how to judge the desirability of a dictionary with 10,000 entries (or with 20,000 entries). 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 find a nondefective dictionary neutral or attractive.

Usually, attributes with dichotomous values (yes versus no), such as whether a dictionary is defective or not or whether an accountant has a CPA license or

not, are easy to evaluate independently because people often know that these attributes have only two alternative values and know which value is good and which one is bad.[6] However, easy-to-evaluate attributes do not have to be dichotomous, as will be shown in the other studies to be reported later.

2. Explaining JE–SE Reversals

According to the evaluability hypothesis, the relative impact of the hard-to-evaluate and the easy-to-evaluate attribute changes between the SE and the JE mode. In the SE mode, the hard-to-evaluate attribute has little impact in differentiating the evaluations of the objects; the easy-to-evaluate attribute is the primary determinant. The reason is as follows: Because people do not have a clear idea of how good each value on the hard-to-evaluate attribute is, two things will happen in SE: (1) the values of the two target objects on the hard-to-evaluate attributes will cast similar impressions, and (2) these impressions will be fuzzy and involve a high variance (see Mellers, Richards, and Birnbaum 1992 for an analysis of the second point). In this circumstance, the hard-to-evaluate attribute will have little power to differentiate the evaluation of one object from the evaluation of the other object. For instance, suppose that participants in the SE conditions of the dictionary study had been asked to rate their impression of the comprehensiveness of the dictionaries on a 0–10 point scale on which increasing numbers indicated more comprehensiveness. Two things would probably have happened: (1) the mean rating given by subjects evaluating only the 10,000-entry dictionary would be similar to that by subjects evaluating only the 20,000-entry dictionary, and (2) the ratings given by either group of subjects would entail a large variance. As a consequence, the entry attribute would not contribute much to differentiating the final valuation of one dictionary from the final valuation of the other dictionary.

In contrast, the easy-to-evaluate attribute would cast different impressions in SE. For example, people evaluating the defective dictionary would have a negative impression of its appearance, and those evaluating the nondefective dictionary would have a neutral or positive impression. Thus, this attribute would discriminate the valuation of one dictionary from the valuation of the other dictionary.

In JE, people could compare one object against the other. This comparison would increase the evaluability of the otherwise hard-to-evaluate attribute and thereby increase its impact on the evaluations of the objects. For example, in the JE condition of the dictionary study, respondents could recognize, through a comparison of the two dictionaries, that the 20,000-entry dictionary

[6] Strictly speaking, the defect attribute in the dictionary study can also be construed as a continuous variable because there are *different degrees* of defectiveness. However, most people would probably be first concerned about whether or not a dictionary is defective before they would care about the degree of defectiveness. Therefore the defect attribute can be construed as a dichotomous variable.

was more comprehensive than the 10,000-entry dictionary. Thus, the number-of-entries attribute, which had little impact in SE, would now help differentiate the valuations of the dictionaries. Joint evaluation may also increase the impact of the easy-to-evaluate attribute. However, because the easy-to-evaluate attribute already has an impact in SE, it will not benefit as much from JE as the hard-to-evaluate attribute. In other words, the hard-to-evaluate attribute has a greater relative impact in JE than in SE, and the easy-to-evaluate attribute has a greater relative impact in SE than in JE. Indeed, the reversal found in the dictionary study supports this analysis.

The evaluability hypothesis is also consistent with the finding of the political candidate study by Lowenthal (1993). In that study, the two candidates varied on two attributes: number of jobs a candidate could bring to the district and whether or not the candidate had a prior conviction. Without a basis for comparison, it was rather difficult to know how good a candidate was if he could bring 1,000 jobs (or 5,000 jobs); in contrast, a prior conviction was obviously undesirable, and a clean history was good. According to the evaluability hypothesis, the job attribute would loom larger in JE, and the conviction attribute would loom larger in SE. The result was indeed in line with this prediction. Likewise, in the dispute resolution study by Bazerman et al. (1992), the two outcomes mentioned above can be interpreted as varying on two attributes: payoff to oneself and whether the payoffs were equal between the two parties. The payoff attribute was relatively difficult to evaluate independently, whereas the equality attribute was easy to evaluate. Again, the reversal observed in that study was consistent with the evaluability hypothesis, implying that the payoff attribute had a greater impact in JE and the equality attribute a greater impact in SE (see Bazerman, Tembrunsel, and Wade-Benzoni, 1988 for an alternative explanation of this study).

IV. EMPIRICAL SUPPORT FOR THE EVALUABILITY HYPOTHESIS

This section presents two studies that tested the evaluability hypothesis: the programmer and the CD-changer studies. The programmer study included naturally occurring, hard-to-evaluate and naturally occurring, easy-to-evaluate attributes. In the CD-changer study, whether an attribute was hard or easy to evaluate was manipulated empirically.

1. The Programmer Study

This study is a replication of the dictionary study with two extensions. First, in the dictionary study (and the other studies reviewed previously), the easy-to-evaluate attribute was always a dichotomous variable (e.g., whether the cosmetic condition of a dictionary was perfect or imperfect), and the hard-to-evaluate attribute was always a continuous variable (e.g., the number of entries

in a dictionary). In the programmer study, both the easy-to-evaluate and the hard-to-evaluate attributes were continuous variables. Second, in the programmer study, a manipulation check was employed to ensure that the attribute believed to be difficult to evaluate was indeed more difficult to evaluate than the attribute believed to be easier to evaluate.

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

	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

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

Note that the two candidates had a trade-off between GPA and experience. Both are continuous variables. For ease of discussion later, let us simplify the descriptions of the two candidates as follows:

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

Respondents (112 students from the University of Illinois at Chicago) were assigned to one of three conditions – JE, SE–A and SE–B. In all three conditions, 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 JE condition, participants evaluated both candidates. In each SE condition, they evaluated only one of the candidates. The evaluation scale was constant across the three versions: willingness to pay.

To assess the evaluability of the attributes, participants in the two separate-evaluation conditions were asked the following questions after they had indicated their WTP for the candidate: (1) "Do you have any idea how good a GPA of 4.9 (3.0) from UIC is?" and (2) "If someone has written 10 (70) KY programs in the last 2 years, do you have any idea how experienced he or 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

Table 31.2. Mean WTP Values (in $1000) for the Two Candidates in the Programmer Study

Evaluation Mode	Candidate A	Candidate B
Joint	$31.2	$33.2
Separate	$32.7	$26.8

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 on which a greater number indicated greater evaluability.

Results and Discussion. The mean evaluability score for GPA was 3.7, and that for experience was 2.1. The difference was highly significant ($t = 11.79$, $p < .001$). These results established GPA as a relatively easy-to-evaluate attribute and experience a relatively hard-to-evaluate attribute in this experiment.

According to the evaluability hypothesis, the experience attribute would have a greater impact in JE than in SE, and the GPA attribute would have a greater impact in SE than in JE. The results, summarized in Table 31.2, supported this prediction.

A clear reversal occurred between JE and SE. In JE, WTP was greater for the more experienced candidate (Candidate B) ($t = 1.65$, $p < .1$). In SE WTP was higher for the candidate with the higher GPA (Candidate A) ($t = 5.50$, $p < .001$).

2. The CD Changer Study

As mentioned earlier, the evaluability hypothesis asserts that JE–SE reversals occur because one of the attributes involved in the stimulus objects is hard to evaluate independently whereas the other attribute is relatively easy to evaluate. It implies that if both attributes are easy to evaluate independently, then there will be no reversal. If this logic is valid, then a JE–SE reversal can be turned "on" or "off" by varying the relative evaluability of the attributes.

In all of the studies discussed thus far, whether an attribute was hard or easy to evaluate independently was assumed and not manipulated empirically. In the study described below, the evaluability of an attribute was manipulated empirically. This manipulation was designed to test whether a JE–SE reversal can indeed be turned on or off as the evaluability hypothesis predicts. The study included two evaluability conditions: Hard–Easy and Easy–Easy. In the Hard–Easy condition, one of the attributes in the stimulus objects was easy to evaluate independently and the other hard to evaluate independently. In the Easy–Easy condition, both attributes were easy to evaluate independently. It was predicted that a JE–SE reversal would be more likely to exist in the Hard–Easy condition than in the Easy–Easy condition.

Method. This study involved the evaluations of two CD changers (i.e., multiple compact disc players) as follows:

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

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:

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

Respondents (202 students from the University of Illinois at Chicago) were assigned to one of six conditions. These six conditions constituted a 3 (evaluation mode: JE, SE–A and SE–B) × 2 (evaluability: Hard–Easy versus Easy–Easy) factorial design. 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 for both CD changers; in each separate-evaluation condition, for only one of the CD changers.[7]

In the Hard–Easy condition, participants received no other information about either THD or CD-capacity than described previously. In this condition, THD was a hard-to-evaluate attribute, and CD-capacity was a relatively easy-to-evaluate attribute. Most people, although they know that less distortion is better, 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 desirable or not.

In the Easy–Easy condition, participants were provided with information about the effective range of the THD attribute. They were told, "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 SE conditions would have some idea where the given THD rating fell in the range and hence whether the rating was good or bad.

To ensure that the evaluability manipulation was effective, participants in the two SE conditions were asked the following questions after they had indicated

[7] In the JE condition of this study, participants were first asked whether they were willing to pay more for A or for B and then indicated how much they were willing to pay for each. Four participants were excluded because they said that they were willing to pay more for one model but gave a higher WTP price for the other.

their WTP values: "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 SE–A condition, and those in the parentheses were for the SE–B condition.) As in the programmer study, answers to those questions ranged from 1 to 4, and higher numbers indicated greater evaluability.

Results and Discussion. First, the evaluability scores for the two attributes indicate that the evaluability manipulation was successful. Specifically, 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 ($F(1,135) = 9.40$, $p < .01$), indicating that the difference in evaluability between THD and CD-capacity decreased significantly from the Hard–Easy condition to the Easy–Easy condition. Planned comparisons found that evaluability scores for THD were significantly higher in the Easy–Easy condition than in the Hard–Easy condition ($t = 2.92$, $p < .01$), whereas those for CD-capacity stayed virtually the same.

Let us now turn to the main dependent variable: willingness-to-pay for the two CD changers. If the evaluability hypothesis is correct, then a JE–SE reversal was likely to exist in the Easy–Hard condition but not in the Easy–Easy condition. The results, summarized in Table 31.3, confirmed this prediction.

In the Hard–Easy condition, there was an unequivocal JE–SE reversal, and its direction was consistent with the evaluability hypothesis, implying that the hard-to-evaluate attribute (THD) lost its impact from JE to SE and the easy-to-evaluate attribute (CD capacity) gained its impact. Specifically, in JE WTP values were higher for CD Changer B than for CD Changer A ($t = 1.96$, $p < 0.1$), but in SE WTP values were higher for CD Changer A than for CD Changer B ($t = 2.70$, $p < .01$). In the Easy–Easy condition, the reversal disappeared. The WTP values were higher for CD Changer B in both JE ($t = 2.81$, $p < .01$) and SE ($t = 2.92$, $p < .01$).

That manipulating attribute evaluability could affect the existence of a JE–SE reversal lends further support to the evaluability hypothesis. It suggests that what drives this type of reversal is differential evaluability between the attributes.

Table 31.3. Mean WTP Values for the Two CD Changers

Evaluability	Evaluation Mode	CD Changer A	CD Changer B
Hard–Easy	Joint	$204	$228
	Separate	$256	$212
Easy–Easy	Joint	$186	$222
	Separate	$177	$222

3. Remarks

Arguably, all decisions and judgments are made either in JE or in SE or some combination of the two. The present work shows that people's responses to the same set of objects can vary drastically between JE and SE. Compared with responses in JE, responses in SE are influenced more by easy-to-evaluate attributes and less by hard-to-evaluate attributes.

Several qualifications are in order here. First, in each of the studies presented above, the dependent variable in both the JE and the SE conditions was willingness to pay. The reason to utilize such a uniform dependent variable is to demonstrate that JE–SE reversals can occur even if the evaluation scale is held constant. In reality, however, JE and SE often involve different evaluation scales. In particular, JE and SE are often naturally confounded with choice and judgment. A choice task inevitably involves JE, requiring the respondent to compare the alternative options simultaneously. On the other hand, a judgment task often entails SE or some combination of JE and SE, allowing the respondent to evaluate one object at a time. As a result, preference reversals in reality often exist between "joint choice" and "separate judgment," and these reversals can potentially be accounted for by both evaluability, and compatibility.

Second, whether an attribute is easy or difficult to evaluate is not an intrinsic characteristic of the attribute; it depends on how much knowledge decision makers have about the attribute and about its context. For example, total harmonic distortion (THD) is a hard-to-evaluate attribute for most people but would be an easy-to-evaluate attribute for an audiophile (see Coupey, Irwin, and Payne 1998 for a related argument). In addition, a given attribute can be difficult to evaluate within a certain range but easier to evaluate in another range. For example, most people would not have a good idea how desirable a music dictionary is if it has 10,000 entries but would find a music dictionary with only 50 entries clearly undesirable.

Besides the evaluability hypothesis, there are other possible explanations for JE–SE reversals. For example, in almost all of the studies demonstrating a JE–SE reversal, one can argue that the hard-to-evaluate attribute is always "more important" than the other attribute. For example, in the dictionary study, the hard-to-evaluate attribute is number of entries, and arguably it is also the more important attribute. Likewise, in the Hard–Easy condition of the CD changer study, the hard-to-evaluate attribute is sound quality (THD), and arguably it is also the more important attribute. Hence, it is tempting to speculate that it is differential attribute importance, rather than differential attribute evaluability, that drives the JE–SE reversal. However, this speculation is inconsistent with the finding of the CD changer study. In that study, increasing the evaluability of THD was able to eliminate the JE–SE reversal. This manipulation should not have affected the importance of THD. If anything, this manipulation may have only increased the importance of THD and, therefore, should have accentuated, rather than attenuated, the reversal. (One may wonder why in most studies

the hard-to-evaluate attribute is also the more important attribute. This is a design feature required to induce a JE–SE reversal effect. In order to have room for a reversal, it is necessary that the hard-to-evaluate attribute be sufficiently important so that the option superior on this attribute will be favored in JE.)

Bazerman and his colleagues (1998) proposed that attributes that people *want* to consider loom large in SE and attributes that people think they *should* consider loom large in JE. Although this proposition is a possible alternative explanation for some of the JE–SE reversals (e.g., those involving trade-offs between payoffs to oneself and fairness), it does not readily apply to many other JE–SE reversals (e.g., Lowenthal's political candidate study and the CD-changer study) in which it is unclear which attribute is the "should" attribute and which is the "want" attribute. Interested in consumer behavior, Nowlis and Simonson (1997) proposed that what they call "comparative" attributes would receive more weight in purchase choices and what they call "enriched" attributes would receive more weight in purchase intention ratings, and they found evidence consistent with their proposition. However, in the studies discussed in this article, it is difficult to know a priori which attributes are comparative and which are enriched. Furthermore, purchase choice and intention ratings differ not just in evaluation mode, but also in evaluation scale.

Although the evaluability hypothesis is not the only explanation for JE–SE reversals, it is probably the most parsimonious and most consistent explanation for a wide range of JE–SE reversals.

Finally, there are also other types of JE–SE reversals than those reviewed above. For example, Irwin et al. (1993) found that in JE people were willing to pay more for improving the air quality in Denver than for improving a consumer product such as a VCR, but that in SE WTP values were higher for improving the consumer product. Kahneman and Ritov (1994) found that in JE people were willing to contribute more to programs that would save human lives (e.g., farmers with skin cancers) than to programs that would save endangered animals (e.g., dolphins), but that in SE the rank order of the two issues was reversed. Unlike the kinds of stimuli examined in the present article, the options compared in Irwin et al.'s and Kahneman and Ritov's studies were of different categories (e.g., air quality versus consumer products) and shared no common attributes. Explanations for these results require a combination of Norm Theory (Kahneman and Miller 1986) and the evaluability hypothesis and are beyond the scope of this article (see Hsee et al. 1999 for details).

V. IMPLICATIONS OF THE EVALUABILITY HYPOTHESIS BEYOND JE–SE REVERSALS

Although the evaluability hypothesis was proposed to account for JE–SE reversals, it has implications for a much broader range of issues. In this section, I discuss the following issues:

1. Why an objectively inferior option may be favored to an objectively superior option in SE,
2. Why people are often more sensitive to the proportion attribute of an object than to its actual outcome value,
3. Why people's belief about the relative importance of attributes sometimes differs from the actual influence of the attributes, and
4. Why the option people choose during the decision phase does not always yield the best experience during the consumption phase.

Let me mention here that these issues may be open to multiple explanations. My intention is not to rule out other explanations but to provide a new perspective based on the evaluability notion.

1. Less Is Better

The evaluability hypothesis suggests that when objects are evaluated in the SE mode, the rank order of these objects is determined primarily by attributes that are easy to evaluate independently, even if these attributes are not the normatively most important attributes. In the extreme case, it is possible that an objectively inferior object may be judged more favorably than an objectively superior alternative under the SE mode.

This possibility was explored in a series of studies reported in Hsee (1998). One study, for example, involved the evaluation of two hypothetical dinnerware sets in a store having a clearance sale. Set A included 24 pieces: 8 dinner plates, 8 soup bowls, and 8 dessert plates. All of the pieces were in good condition. Set B comprised 40 pieces: 8 dinner plates, 8 soup bowls, 8 dessert plates, 8 cups, and 8 saucers. All of the plates, soup bowls, and dessert plates were in good condition, but two of the cups and seven of the saucers were broken. Note that Set B contained the same 24 intact pieces as Set A, plus it also contained two intact cups and one intact saucer that were not available in Set A. Therefore, Set B should be of a greater value than Set A. However, when these two sets of dinnerware were evaluated separately, people were willing to pay significantly more for Set A, demonstrating a less-is-better effect. The valuations were reversed when the two sets of dinnerware were presented jointly.

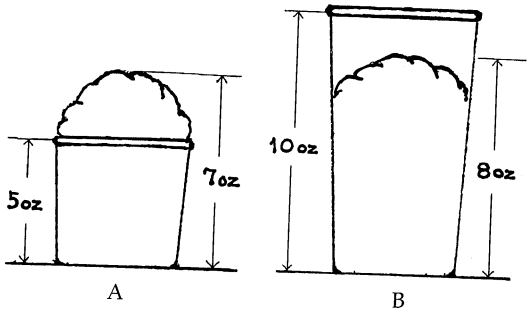
The less-is-better effect revealed in the SE condition of the dinnerware study can be readily explained by the evaluability hypothesis. The two sets of dinnerware can be considered as varying on the following attributes:

	Dinnerware A	**Dinnerware B**
No. of usable pieces	24	31
Integrity of the set	complete	incomplete

In SE, it was easy to know that a complete set was good and a defective set was bad. However, for most respondents (college students), it was more difficult to determine the desirability of a set with 24 (or 31) usable pieces. Thus, the

difference in valuation between the dinnerware sets in SE was determined mainly by the integrity attribute.[8]

Another study reported in Hsee (1998) involved the evaluation of two hypothetical servings of Häagen-Dazs ice cream. Serving A had 7 ounces of ice cream, and it was contained in a 5-ounce cup. Serving B had 8 ounces of ice cream, and it was contained in a 10-ounce cup. Thus, Serving A was overfilled and Serving B was underfilled. These servings were illustrated graphically as follows:



The differences between the two servings can be summarized as follows:

	Serving A	**Serving B**
Size of cup:	5 oz	10 oz
Amount of ice cream:	7 oz	8 oz
Filling:	overfilled	underfilled

Objectively, Serving B was more valuable because it contained more ice cream (and included a bigger cup!). However, when the two servings were evaluated separately, people were willing to pay significantly more for Serving A. The effect was reversed when the two servings were evaluated jointly.

A similar explanation to the dinnerware study applies here. Even though the amount of ice cream is what people should consider when evaluating how much they would pay for a serving, this attribute was difficult to evaluate independently. Without something to compare it with, most people would not know whether a serving with 7 ounces (or 8 ounces) of ice was desirable or not. In contrast, the filling attribute was easy to evaluate independently: People would

[8] In the dinnerware study, respondents may have inferred the quality of a dinnerware set based on whether it contained any broken pieces. For example, those evaluating Set B may have inferred that it must be a low-quality set because some pieces in that set were broken. Although this possibility can only explain why WTP for Set B was lower in SE, it does not explain the JE–SE reversal. If respondents had indeed made such inferences, they should have done so in both joint evaluation and separate evaluation. It should also be noted that in most of the studies demonstrating a JE–SE reversal, the attributes are entirely independent, that is, there is no reason for one to infer the desirability of one attribute from that of the other attribute. For example, in the dictionary study, whether a dictionary had a torn cover or not had nothing to do with how comprehensive the dictionary was.

consider an overfilled cup appealing and an underfilled cup unappealing. Thus, the filling attribute became the primary determinant of willingness-to-pay in SE.

2. Proportion and Base Number

Oftentimes, we need to evaluate outcome values that are expressed as proportions of some base numbers. For example, how good is an environmental protection program that can save 20% of 10,000 endangered wild birds in a certain forest? According to the evaluability hypothesis, in SE of such problems, we are often sensitive to the proportion attribute and insensitive to the base number attribute or to the actual outcome value attribute. The following example is inspired by the recent finding of Fetherstonhaugh et al. (1997) that programs expected to save a given number of lives received greater support if the number of lives at risk was smaller (see also Baron 1997 and Jenni and Loewenstein 1997). The example to be discussed below is about two environmental protection programs, each of which could save a certain number of wild birds in a given forest. In particular,

	Program A	**Program B**
Size of the forest:	has 5,000 birds	has 25,000 birds
Number of birds the program can save:	4,000	5,000
Percentage of birds in the forest the program can save:	80%	20%

Here, the size of the forest is a base number, number of birds the program can save is the actual outcome value, and the outcome value is a proportion of the base number. According to the evaluability hypothesis, when the two programs are evaluated separately, Program A will be favored. The reason for this prediction is simple: The proportion attribute is relatively easy to evaluate independently. It has a natural lower bound (0%) and a natural upper bound (100%), and people would know that 80% is a reasonably high number and 20% a low number. In contrast, both the base number and the outcome value are difficult to evaluate independently because most people would not know the range, distribution, or other reference information of these attributes. As a result, people's responses in SE are often dominated by the proportion attribute. The evaluability hypothesis also predicts that if either the base number or the outcome value is made easier to evaluate independently (e.g., by giving people more reference information), the relative impact of the proportion attribute will diminish (see Wright 1997 for other factors that can influence the relative impact of proportions and absolute differences).

The previous analysis can also explain why speakers are usually happier with the number of people who attend their talk if they are assigned a small room (say with 50 seats) and a high proportion of the seats (say 80%) are occupied than if they are assigned a larger room (say with 250 seats) and only a small

proportion (say 20%) of the seats are occupied even though the talk is better attended in the latter scenario. The logic of this example is identical to that of the bird example.

A closer look suggests that the structure of the preceding examples is also parallel to the structure of the ice cream study discussed earlier. The size of the forest or the size of the room is like the size of the cup in the ice cream study. The number of birds the program can save or the size of the audience is like the amount of ice cream in a serving; these are the actual outcome values. The proportion attribute is like the filling attribute in the ice cream study and reflects the relationship between the base number and the outcome variable. In all of these cases, the option with the inferior outcome value is favored in SE.

3. Believed Importance versus Actual Influence[9]

When evaluating an object involving multiple attributes, we may ask ourselves about the relative importance of the attributes. For example, when evaluating a candidate, we may wonder whether his or her appearance or experience is more important. Presumably, our belief about the relative importance of the attributes should be consistent with how much influence these attributes actually have on our evaluation. However, this may not always be the case. Sometimes, we may believe that a certain attribute is more important than another (e.g., experience is more important than appearance), but the attribute believed to be more important may be hard to evaluate and therefore have little influence on our evaluation, especially when the evaluation is made in SE.

The preceding intuition was tested in a preliminary study involving the evaluation of a hypothetical job candidate for a programmer position. The study included 4 conditions, among which the candidate varied on 2 attributes: (1) experience (the candidate had written either 110 programs or 220 programs), and (2) appearance (he was either unkempt and messy or well-groomed and nicely dressed). These 2×2 conditions were presented between subjects so that the evaluation mode was always SE. Note that here, experience was more difficult to evaluate independently than appearance.

Respondents were asked three questions. The first asked what salary they would be willing to pay should the candidate be hired. The second and third questions assessed their belief about the relative influence of the attributes. The second question asked whether their salary decision in the first question was influenced more by the candidate's experience or by his appearance. The third question provided range information about the attributes and asked which of the following would make a greater difference in their salary decisions: (a) whether the candidate had written 110 or 220 programs, or (b) whether the candidate was unkempt and messy or well-groomed and nicely dressed. The reason to include both the second and the third questions was to ensure

[9] The idea presented in this section was inspired by discussions with Paul Slovic. The appearance–experience experiment described later is an unpublished study we designed to test the idea.

reliability. Previous research suggests that judgment of relative importance may change, depending on whether range information is given (e.g., Goldstein 1990, Goldstein and Mitzel 1992).

The results revealed a sharp inconsistency between actual influence and beliefs. In the actual WTP decisions, experience had virtually no effect, and appearance had a significant effect. People were willing to pay approximately $1,800 more for the well-groomed and nicely dressed candidate than for the messy and unkempt one. However, when asked which attribute had a greater influence on their WTP, 85% of the respondents said it was experience. Even when the ranges of the attributes were given (the last question), still the majority (84%) believed that experience would make a greater difference than appearance. These beliefs were grossly inconsistent with the actual influence of the attributes.

This study suggests that what people believe to be important for an evaluation may be at odds with what actually affects the evaluation. This inconsistency is most likely to occur if the evaluation task is made in SE and the attributes believed to be important are difficult to evaluate independently. In situations like this, the presumably important attributes will have little or no impact on the actual evaluation.[10]

4. Inconsistencies between Decision and Its Consequence

The most speculative, but also potentially the most important, implication of the present research has to do with discrepancies between people's decisions and their subsequent experience with the consequences (Kahneman and Snell 1990, 1992). Ideally, when making a decision, people ought to choose the option they will like when they use it or "consume" it later. In reality, people often fail to do so. For example, when making a decision about which piano to purchase, people may choose one model over another model, but after they purchase the piano and move it home, they may not like it. There are many possible reasons for such decision–consumption inconsistencies, and there is ample research on this topic. For example, people may not have as much information about the options at the decision phase as at the consumption phase, people may be at different arousal states (hot versus cold) during the two phases (Loewenstein 1996), or people's taste may change over time (e.g., Kahneman and Snell 1992, March 1978), to name just a few.

However, there is an important but largely neglected contributor to decision–consumption inconsistency: evaluation mode. More often than not, a decision is made in the joint evaluation mode, and the consequence of a decision is experienced or consumed in the SE mode. For example, when a person buys

[10] Another common reason why believed importance may differ from actual influence is that the attribute believed to be important has no variance in its values and therefore has no effect. In the study described above, however, the reason why experience had no effect was not that it had no variance in its values but that the variance in its values (i.e., 110 programs versus 220 programs) failed to generate a corresponding variance in the *evaluations* of these values because the attribute was difficult to evaluate independently.

a piano in a musical instrument store, there are typically myriad models for her to compare and choose from (JE). However, after she buys a piano and when she "consumes" it at home, that is, plays it, looks at etc., she is exposed mostly to that particular piano alone (SE). (Of course, she may also occasionally think about the forgone alternative during the consumption phase, but usually the evaluation mode in the consumption phase, though not purely SE, is much closer to SE than the evaluation mode in the decision phase.) Likewise, when we decide which candidate to vote for during a presidential election, we are in the JE mode, comparing one candidate with another. Once a candidate becomes the president and starts his or her term, we find ourselves mostly in the SE mode, facing that particular president alone. Shafir (in press) argues that the distinction between joint and separate evaluation has even wider implications. He proposes that guidelines and policies are borne out of joint evaluation of alternative scenarios, but events in the real world, to which these guidelines and policies are supposed to apply, usually present themselves one at a time.

To the extent that decisions are made in JE and consumption takes place in SE, reversals between JE and SE will manifest themselves between decision and consumption. Generally speaking, hard-to-evaluate attributes loom larger in decision making, and easy-to-evaluate attributes loom larger in consumption experience.

The analysis presented above implies that decision makers may overpredict the impact of hard-to-evaluate attributes on their consumption experience. As discussed earlier, hard-to-evaluate attributes have little discriminatory power in separate evaluation and hence make little differences to one's consumption experience in separate evaluation. However, people may not realize this when they make decisions in joint evaluation and may place too much weight on these attributes. The following story illustrates this point.

Years ago, somebody (call him Mr. S) shopping for a pair of speakers in an audio store was ushered by a salesperson to a soundproof listening room in which one could easily compare different models of speakers. After a few minutes, Mr. S narrowed his choices to two models. They had the same price, but one looked attractive and the other ugly. He then played his favorite CD on the two models, and, after careful comparisons, found the ugly-looking model sounded slightly better. Thinking that sound is important for speakers, he made the decision to buy the better-sounding model. After he took the speakers home, he was happy with them in a honeymoon sort of mood for only a few days and then became increasingly annoyed by their appearance. Before long, he relegated these speakers to the basement and has never bothered to listen to them again. Although there is no way to find out if it is true, my speculation is that he would have been happier if he had bought the other model. Chances are that, unless through direct comparisons, as in the listening room, he could not even tell the difference in sound quality between the two models. However, the model he did not buy apparently looked attractive and would have made him happier.

For most consumers, the sound quality of a speaker, at least within a certain range, is hard to evaluate independently, and the difference in sound quality between two models of speakers can only be appreciated in JE, not in SE. In contrast, the look of a speaker is much easier to evaluate independently. The moral of this story is: When making decisions, people put too much weight on hard-to-evaluate attributes and are too obsessed with differences between options that will make little difference in SE and hence little difference in consumption experience.

VI. CONCLUSION

When two objects involving a trade-off between a hard-to-evaluate attribute and an easy-to-evaluate attribute are evaluated, the evaluations of these objects may change dramatically, depending on whether the evaluations are elicited in JE or in SE. A natural question to ask here is, which evaluation mode is better, JE or SE? The answer depends on the purpose of the evaluation. If the purpose is to be consistent with the objective quality of the target objects, then JE is better. As previous examples illustrate, SE may lead to a higher valuation of an objectively inferior option. That is probably why decision experts (e.g., Janis and Mann 1977) usually advise people to compare alternatives and engage in JE when making decisions.

However, even in decision making, JE is not always better than SE. If the decision maker's purpose is to choose the option he or she will enjoy later, then SE may be better. In Mr. S's story, for instance, if his purpose was to buy the speakers that he would like after bringing them home, then he should have engaged in SE when deciding which model to purchase. In reality, however, it is impossible to engage in pure SE when faced with multiple options, but it is possible to achieve something close to SE. For instance, what Mr. S should have done was not to compare the alternatives side by side. Instead, he should have studied the two models on two separate days, focused on one model each day, imagined how he would like it at home, and written down his overall impression. He should then have bought the model with the better overall impression score.

It is clear from the current research that responses made in JE can be dramatically different from those made in SE and that the differences are predictable. But it is not clear whether JE or SE is better. In the last sections of this article, I provided evidence showing that SE can lead to preferences for objectively inferior options. At the same time, I also presented examples and speculations showing that decisions made in SE may lead to better future consumption experience. The reader is the ultimate judge about which evaluation mode is better.