Options as Information: Rational Reversals of Evaluation and Preference

Shlomi Sher and Craig R. M. McKenzie
University of California, San Diego

This article develops a rational analysis of an important class of apparent preference reversals—joint-separate reversals traditionally explained by the evaluability hypothesis. The "options-as-information" model considers a hypothetical rational actor with limited knowledge about the market distribution of a stimulus attribute. The actor's evaluations are formed via a 2-stage process—an inferential stage in which beliefs are updated on the basis of the sample of options received, followed by an assessment stage in which options are evaluated in light of these updated beliefs. This process generates joint-separate reversals in standard experimental designs. The normative model explains why the evaluability hypothesis works when it does, identifies boundary conditions for the hypothesis, and clarifies some common misconceptions about these effects. In particular, it implies that joint-separate reversals are not irrational; in fact, they are not preference reversals. However, in expanded designs where more than 2 options are jointly evaluated, the model predicts that genuine (and rational) preference reversals will sometimes emerge. Results of 3 experiments suggest an excellent fit between the rational actor model and the judgments of human actors in joint-separate experiments.

Keywords: context effect, evaluability hypothesis, preference reversal, rationality

Procedure invariance is viewed as a basic requirement of rational choice. This principle states that different methods of eliciting preferences from a decision maker (DM) should yield the same ordering of options. However, several apparent violations of procedure invariance have been documented (reviewed in Hsee, Zhang, & Chen, 2004). This article focuses on one influential class of apparent preference reversals—joint-separate reversals (JSRs) traditionally attributed to the evaluability hypothesis (Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999). These effects are often taken to exemplify the divergence between normative and descriptive models of decision making (Hsee et al., 2004; MacLean, 2002; Shafir & LeBoeuf, 2004).

In this article, we develop a normative "options-as-information" model of these JSRs. The model considers a hypothetical rational actor in a joint-separate experiment. The actor is rational but not omniscient. His or her beliefs are updated as new options are sampled, and evaluations are updated in light of these inferences. The model implies that JSRs are not preference reversals and are not counter-normative. Instead, when the dynamically evolving belief states of the DM are accounted for, typical JSRs are seen to exemplify the convergence of normative and descriptive analysis, and to illustrate how these levels of analysis can be mutually

illuminating (Anderson, 1991; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McKenzie, 2003, 2005; Oaksford & Chater, 1994).

Joint-Separate Reversals

In a paradigm developed by Hsee and colleagues (Hsee, 1996, 1998, 2000; Hsee et al., 1999), two options are defined on two critical attributes. One attribute is highly pertinent but "difficult to evaluate"—i.e., participants have minimal prior knowledge of what counts as a good or bad value on the attribute dimension. The other attribute is somewhat less important but far easier to evaluate. The options trade off, so that one option is superior on each attribute. Some participants (in two *separate evaluation* conditions) state willingness-to-pay (WTP) for a lone option, while other participants (in a *joint evaluation* condition) state WTP for both options simultaneously. The general finding is that the option that is superior on the difficult-to-evaluate attribute receives the lower separate evaluation and the higher joint evaluation.

For example, in a study reported by Hsee (1996), participants imagined that they were hiring a programmer who would be using a programming language called KY and who would receive a salary between $20,000 and $40,000. Candidates differed on the easy-to-evaluate attribute of grade point average (GPA) and the difficult-to-evaluate attribute of KY programming experience, as seen in Table 1. Participants in this study knew that GPA at University of Illinois Chicago (UIC) is scored on a 5-point scale. In this study, A (who is best on the easy-to-evaluate attribute) received a higher salary than B when candidates were considered alone, but B received a higher salary in joint evaluation. Note that it is unclear how good 10 or 70 KY programs is when a candidate is seen in isolation, but B clearly has more experience than A when the candidates are seen jointly. Similar effects have been demonstrated in the domains of consumer choice (Hsee, 1996), medical

Table 1
Programmer Problem (Hsee, 1996)

	Candidate A	Candidate B
Education	BS in computer science from UIC	BS 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

Note. BS = bachelor of science; GPA = grade point average; UIC = University of Illinois Chicago.

decision making (Zikmund-Fisher, Fagerlin, & Ubel, 2004), and political attitudes (Lowenthal, 1993, cited in Hsee et al., 1999).

The favored explanation for this phenomenon is the evaluability hypothesis (Hsee, 1996; Hsee et al., 1999). This hypothesis distinguishes two "modes" of evaluation—separate evaluation (SE) mode, in which a single item is evaluated, and joint evaluation (JE) mode, in which multiple options are assessed. The evaluability hypothesis states that, when attributes differ in ease of evaluation, "the relative impact of the difficult-to-evaluate attribute increases from SE to JE" (Hsee et al., 1999, p. 581). Consistent with the hypothesis, JSRs disappear when distributional information is provided, making the otherwise difficult-to-evaluate attribute more readily interpretable (Hsee, 1996).

For the discussion that follows, it will be useful to express the evaluability hypothesis in formal terms. In general, for an option *a* and an option-set *V* containing it, let $WTP(a|V)$ denote the participant's evaluation of *a* when option-set *V* is received. In the joint-separate design, let *d* be the option that is superior on the difficult-to-evaluate attribute, and let *e* be the option that is superior on the easy-to-evaluate attribute. We consider a general joint-separate design, in which three option-sets are presented for evaluation in three experimental conditions—the singleton sets $\{d\}$ and $\{e\}$ and a joint-evaluation set *U* with $\{d, e\} \subseteq U$. Most joint-separate experiments target the simplest version of this design, in which $U = \{d, e\}$. In analyzing a joint-separate experiment, we are interested in the relationship between the following two WTP-differences:

$$\Delta_{sep} = WTP(d|\{d\}) - WTP(e|\{e\})$$

$$\Delta_{joint} = WTP(d|U) - WTP(e|U)$$

Δ_{sep} is the observed difference between separate evaluations, and Δ_{joint} is the difference between joint evaluations. Note that both differences are signed, with larger positive values corresponding to a larger advantage for *d* over *e*. The evaluability hypothesis implies the following general relation, which we call the *Evaluability Inequality* (EI):

$$\Delta_{joint} > \Delta_{sep}$$

The JSR is the special case of the EI in which $\Delta_{joint} > 0 > \Delta_{sep}$. Note that the evaluability hypothesis does not predict that JSRs will always occur. Rather, the hypothesis entails that the EI generally obtains, and this explains why JSRs sometimes occur. Recently, Hsee and Zhang (2010) introduced *general evaluability theory* (GET), which generalizes and extends the evaluability

hypothesis and which we examine in the General Discussion. Importantly, GET also implies the EI.

JSRs are invariably classed as preference reversals. As preference reversals, these effects are said to violate procedure invariance and are generally deemed counter-normative. For example, in reviewing JSRs and two other preference reversal paradigms, Hsee et al. (2004) wrote, "All these preference-reversals are violations of procedure invariance . . . an important pillar of rational choice" (p. 361) These authors see preference reversals as "a prime example of internal inconsistency" (p. 360) in decision making. More specifically, the evaluability hypothesis has been taken to imply that "the impact an attribute has on one's judgment may not be proportional to the normative importance of the attribute, but rather to its evaluability" (Hsee, 1998, p. 118). The apparent underweighting of difficult-to-evaluate attributes in separate evaluation has been deemed a "systematic cognitive bias" (Bazerman, Moore, Tenbrunsel, Wade-Benzoni, & Blount, 1999, p. 51).

In the following pages, we propose a rational analysis of JSRs. This analysis explains why the evaluability hypothesis works in standard designs. It also identifies boundary conditions that limit its scope, and it clarifies the normative status of these effects. The model has four main implications: First, standard JSRs are not counter-normative. Second, they are not preference reversals. Third, while the normative model and the evaluability hypothesis both predict the EI in two-option designs (where $U = \{d, e\}$), the normative model implies that the EI should become unstable in expanded designs where more than two options are jointly evaluated. Finally, the model predicts that, in such expanded designs, genuine—and rational—preference reversals should sometimes occur. We report three experiments that suggest an excellent fit between the rational actor model and the judgments of human actors in joint-separate experiments.

While we believe the normative model provides a compelling account of JSRs where a single attribute is unfamiliar, the present analysis does not apply to other joint-separate effects (e.g., Irwin, Slovic, Lichtenstein, & McClelland, 1993; Kahneman & Ritov, 1994) in which asymmetric unfamiliarity of attributes is not critical. These effects, along with response mode reversals and some widely studied context effects, are revisited in the General Discussion.

Options as Information

Models of rational choice assume stable preferences in a fixed information state. However, when prior knowledge of the stimulus universe is limited, the presentation of options alters the information state of the DM, and hence may alter preferences and evaluations. Figure 1 depicts the "rational actor" model that is appropriate in this case.

The model involves a two-stage process of inference and evaluation. The first stage moves from options to inferences. The DM begins with an underspecified model M, which captures his or her limited initial beliefs about the universe of options. A set $\{o_i\}$ of options is then received. This option set may be regarded as a sample drawn (not necessarily at random) from a natural population of options (i.e., "the market"). On obtaining this sample, the DM updates his or her beliefs to arrive at an enriched posterior model M^* of the option space.
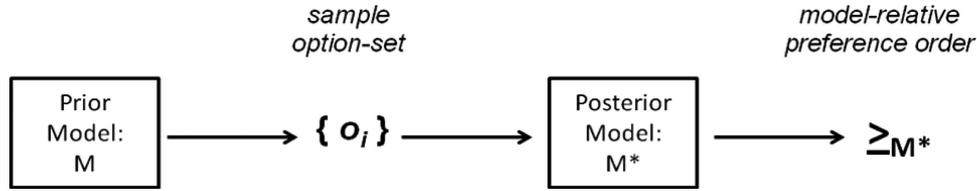
Figure 1. Normative options-as-information model for a generic decision making experiment. The prior model M embodies the decision maker's beliefs after reading the background information but before seeing specific options. $\{o_i\}$ is the set of options encountered in the choice phase of the experiment. The posterior model M* reflects the inferences drawn from this option-sample. This updated "view of the market" induces a preference order $\geq_{M^*}$ over the options that compose it.

The second stage moves from inferences to evaluations: The updated model induces an updated preference order $\geq_{M^*}$, which captures the relative attractiveness of options in light of the DM's posterior beliefs about the market. Importantly, different samples may lead to different posterior models M* and hence to different preference orders $\geq_{M^*}$. As a result, rational evaluations will be coherent within but not across contexts.

The joint-separate design features one attribute (e.g., GPA) for which prior knowledge is maximal, and hence for which sample-based inference will be minimal. However, prior knowledge is minimal for a second attribute (e.g., programming experience), and sample-based inference may be substantial for this attribute. The initial updating stage will thus be characterized by a normative *assimilation effect* for the difficult-to-evaluate attribute, in which the parameters of the posterior model are drawn toward the statistics of the option-sample. For example, rational DMs will infer a higher average number of programs written by candidates on the market when the evaluation set is {A, B} (sample mean of 40) than when the evaluation set is {A} (sample mean of 10), and the estimated average will be still higher when the evaluation set is {B} (sample mean of 70). As we discuss below, the size of the normative assimilation effect depends on the DM's assumptions about the experimental environment.

Whereas the first stage of the rational evaluation process is characterized by an assimilation effect, the second stage, moving from inferences to evaluations, gives rise to a normative *contrast effect*. In the typical joint-separate experiment, participants are instructed to imagine that they intend to pay between $x and $y for a product of some kind. The natural construal of this task is that the best options on the market should receive a valuation close to $y, while the poorest options receive a valuation near $x. What matters to evaluation, then, is an option's *relative* standing in the DM's model of the market. Therefore, as the attribute distribution in the DM's posterior model shifts downward, the valuation of any fixed option should correspondingly increase. The worse the market looks, the better any fixed option comes out by contrast.

For concreteness, we consider a DM who consistently employs a specific evaluation rule. For each attribute, values are first normalized to the distribution in the DM's posterior model (e.g., transformed to percentiles or Z-scores). WTP is then computed as a weighted sum of normalized attribute values. Importantly, we assume that the weight attached to relative standing on each attribute is the same in all contexts. Weighted-additive (WADD) rules figure prominently in normative models of multiattribute choice (Payne, Bettman, & Johnson, 1993). We employ a relative

WADD (rWADD) rule because, as noted above, the joint-separate task requires an assessment of relative standing.

Figure 2 illustrates how the normative process of inference and evaluation can lead to a JSR in the programmer problem. The middle column depicts the posterior models of programming experience inferred when each evaluation set is sampled. For illustrative purposes, we assume strong assimilation for this attribute. The posterior GPA distribution is assumed to be fixed across conditions and is not shown. The right column depicts the pattern of valuations that results under the rWADD rule.

For each single-option sample, the sampled option is roughly average relative to the resulting posterior distribution of experience. A's experience, normalized to the model inferred when {A} is sampled, is about equal to B's experience, normalized to the
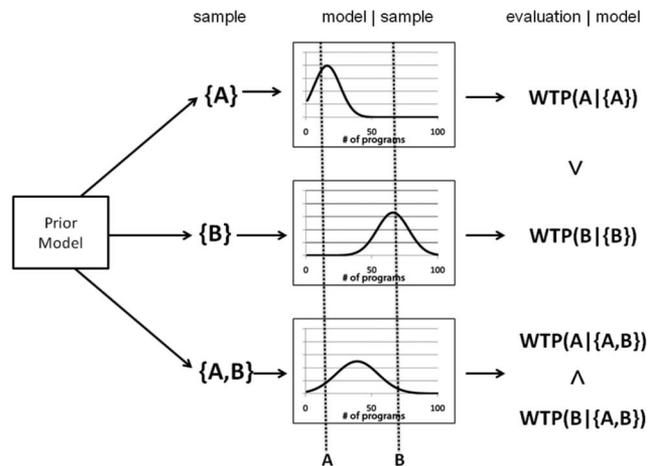
Figure 2. Options-as-information model applied to the programmer problem. The three option-samples lead to distinct posterior models ("model | sample") of the difficult-to-evaluate attribute (programming experience). The posterior distribution of the easy-to-evaluate attribute (grade point average [GPA]) is the same in all conditions and is not shown. Decision weights are assumed to be constant in all conditions, with normalized experience receiving a consistently large weight. The top WTP-inequality reflects the negligible difference between A and B in programming experience when numerical values are normalized to the relevant posterior models; however, normalized GPA favors Candidate A. The lower WTP-inequality reflects the advantage for B in programming experience relative to the single posterior model inferred when {A, B} is sampled. WTP = willingness-to-pay.

model inferred when {B} is sampled. However, normalized GPA is defined relative to a fixed distribution, and always favors A over B. When normalized experience and GPA enter into a weighted sum, the result is that $WTP(A|\{A\}) > WTP(B|\{B\})$.

However, when {A, B} is sampled, both candidates are evaluated relative to a single posterior model of experience, in which A is likely to emerge as below and B as above average. The relation between $WTP(A|\{A,B\})$ and $WTP(B|\{A,B\})$ will then reflect a trade-off between the greater normalized experience of B and the greater normalized GPA of A. If the fixed weight attached to normalized experience is sufficiently large, $WTP(B|\{A,B\}) > WTP(A|\{A,B\})$. That is, a JSR will result.

The magnitude of the normative assimilation effect depends on the DM's conception of the process whereby option samples are selected. For example, the DM may believe that sampling is representative—i.e., that option sets are assembled so that salient statistics approximate the corresponding parameters of the market distribution. Representative sampling may be a natural assumption in a social setting where a knowledgeable party (the experimenter) asks a second party (the participant) about a situation of interest to the former (Hilton, 1995). Under this assumption, the assimilation effect will be large even if samples are small. Alternatively, the DM may assume that options are sampled independently and at random from the market. The magnitude of normative assimilation under a random sampling assumption will depend on the fine details of the DM's prior beliefs and task strategy.[1] Absent a detailed specification of the DM's assumptions, the magnitude of the assimilation effect cannot be stipulated a priori in a normative analysis. However, it can be empirically measured.

The options-as-information model provides a critical new test of coherence in the joint-separate design. We can empirically assess, first, the inferences that are drawn from different samples; and second, the valuations that are formed in light of different beliefs. The model predicts that the observed valuation pattern in a JSR should correspond to the product of (a) the effect of samples on models and (b) the effect of models on WTP.

A thought experiment may illustrate the point. Consider a DM who sees one or two job candidates, and draws private inferences about the distribution of programming experience. Now imagine that we could transplant the posterior model inferred by this DM into the mind of a second DM. In light of the transplanted model, the transplantee evaluates a *single* candidate. Thus the "mode" of evaluation is always separate, and the overall quantity of information that transplantees have is fixed. According to the options-as-information model:

1. The separate evaluation of A by a DM transplanted with a model inferred from sampling {A} should exceed the separate evaluation of B by a DM transplanted with a model inferred from sampling {B}.

2. The separate evaluation of B by a DM transplanted with a model inferred from sampling {A, B} should exceed the separate evaluation of A by a DM transplanted with the same model.

Of course, we have no way of literally transplanting one DM's view of the world into the mind of another DM. Nonetheless, in Experiment 1 we attempted to approximate this procedure in a yoked design.

Experiment 1

The options-as-information model makes two central predictions. First, the different evaluation sets in a joint-separate design lead to different inferences. Second, these different inferences lead to the different evaluations in a JSR. Experiment 1 used the programmer problem (Hsee, 1996) to test these predictions in a yoked design. "Modeler" participants evaluated one or both job candidates, and then reported their beliefs about the distribution of programming experience. The reported beliefs of modeler participants were then provided as background information to "recipient" participants, each of whom evaluated only one candidate. We could thus examine separate evaluations based on posterior models inferred from different option-samples.

When participants are exposed to extensive samples of unfamiliar attribute values, they appear to update their interpretation of the attribute scale in reasonable ways (Beattie & Baron, 1991; Mellers & Cooke, 1994). The modeler component of Experiment 1 tests the prediction that participants will also draw large inferences from small experimental option-samples. This prediction is motivated by prior findings that participants draw robust inferences from subtle features of the experimental environment, such as the framing of options (McKenzie & Nelson, 2003; Sher & McKenzie, 2006) and the wording of questions (Schwarz, 1999) in choice experiments and opinion surveys. It is also compatible with research in social psychology which depicts the participant as an avid information scavenger, drawing far-reaching inferences from fragmentary stimuli (Ambady & Rosenthal, 1993) and subtle pragmatic cues (Hilton, 1995).

If participants draw substantially different inferences from different evaluation sets, the critical question addressed in Experiment 1 is whether these inferences suffice to reproduce the valuation pattern in a typical JSR.

Method

Two hundred sixty-nine undergraduate students at the University of California, San Diego (UCSD) participated in Experiment

---

[1] If random sampling is expected, the inferences supported by small samples depend on the detailed assumptions and strategy of the DM. Under certain conditions, these inferences will be large. For example, suppose the DM can place bounds on the variance of the empirical distribution but has little idea of the mean. Then even a single random observation may lead to a dramatic collapse of a diffuse prior distribution of possible values about a narrow interval. This is because a tight prior bound on empirical variance will constrain the posterior distribution to a small neighborhood about the single observed value. When the DM has little idea of mean or variance, the magnitude of normative assimilation may depend on subtle aspects of the DM's task-strategy. In the main text, we gloss over the fact that the posterior model M* may have either a *modal* or an *average* form. In the former case, the model captures the DM's "best guess" about how the market really looks. In the latter case, the model's distribution captures the DM's uncertainty about how the market might look and is computed as a sum of possible market distributions weighted by their posterior probability. Whether an average or a modal posterior model is appropriate depends on whether the DM wants to minimize expected error or maximize the probability of an accurate response. Both modal and average posterior models will be subject to assimilation, but the effect will generally be more dramatic in the modal case.

1 for partial credit in psychology courses. For all three experiments reported here, the experiment was the first in a series of unrelated studies lasting less than an hour. Experiments 1 and 2 were run on MATLAB using the Psychophysics Toolbox (Version 2.54; Brainard, 1997). Participants were run at individual computer stations, separated by partitions, in groups of up to six. Data were not obtained from two participants in Experiment 1 because of computer errors.

We first describe the procedure for modeler participants. After an introductory screen, modelers were instructed to

> Imagine that you are the owner of a consulting firm, and you are looking to hire a new programmer. The programmer you hire will have to use a special programming language called YT. You are planning to pay your new programmer between $30,000 and $60,000 a year.

A table below this blurb presented information about one or two job candidates. When two candidates were seen, the information was displayed as seen in Table 2. Note that GPA at UCSD is scored on a 4-point scale. This problem is adapted with minor changes from Study 2 of Hsee (1996). When a single candidate was seen, the column displaying the attribute values was simply headed "Candidate." Each modeler was assigned to one of three conditions: *{A}-modelers* saw only the candidate with the higher GPA ($n = 41$); *{B}-modelers* saw only the candidate with more YT experience ($n = 37$); and *{A,B}-modelers* saw both candidates ($n = 40$).

When the participant pressed the spacebar, one or two WTP questions appeared below the blurb and table. For each candidate in the table, the participant was instructed to type a number between 30 and 60 to indicate how many thousands of dollars they would be willing to pay the candidate. The WTP question(s) then disappeared and were replaced by a YT-estimation question, which asked, "How much experience do you think typical candidates for this job have?" (The blurb and candidate table remained visible above this question.) Participants responded by typing numbers into empty text boxes to complete the following sentence: "Typical candidates for this job have written somewhere between _____ and _____ programs in YT, with an average of _____ YT programs." Participants were instructed to make their best guess if they were unsure.

Modelers were included in the yoked design only if their YT estimates were coherent—i.e., the estimated low YT value had to be less than or equal to the estimated average YT value, which, in turn, had to be less than or equal to the estimated high YT value. Coherent YT estimates were provided as background information to one or two recipient participants, according to the following procedure.

Table 2
Stimuli for Experiment 1

	Candidate A	Candidate B
Education	BS in computer science from UCSD	BS in computer science from UCSD
GPA	3.8	3.1
Experience with YT	has written 10 YT programs	has written 70 YT programs

Note. BS = bachelor of science; GPA = grade point average; UCSD = University of California, San Diego.

For recipient participants, a new line with a further sentence was appended to the general background blurb above. For some numbers $x < y < z$, this sentence read, "Typical candidates for this job have written somewhere between x and z programs in YT, with an average of y YT programs." Unbeknownst to the recipient, the values of x, y, and z were equal to the low, average, and high YT estimates of a prior modeler participant. Below this expanded blurb, a table provided information about a single candidate (simply labeled "Candidate"). When the participant pressed the spacebar, a single question below the table requested the participant's WTP for this candidate.

Recipients were yoked to modelers according to the following design: Each modeler's estimate was provided as background information to the next one or two participants who were run on the same computer. Each {A}-modeler's estimate was provided as background information to a recipient who evaluated A alone. Each {B}-modeler's estimate was provided to a recipient who evaluated B alone. Each {A,B}-modeler's estimate was provided to *two* recipients—one who evaluated A alone, and one who evaluated B alone. Letting $X|\{ \cdots \}$ denote the condition in which option X is evaluated based on the YT estimates of a modeler who sampled set $\{ \cdots \}$, there were thus four groups of recipients in Experiment 1: $A|\{A\}$-recipients ($n = 36$), $B|\{B\}$-recipients ($n = 37$), $A|\{A,B\}$-recipients ($n = 38$), and $B|\{A,B\}$-recipients ($n = 38$).

In responding to WTP questions, participants occasionally provided numbers outside the range (30–60 thousand) specified in the instructions. In these cases, the direction of deviation from the range may be informative, but the inclusion of raw responses may introduce outliers. Accordingly, in all experiments, we treated responses below the range as equal to the minimum and responses above the range as equal to the maximum of the instructed range. This affected two WTP responses in Experiment 1.

Results

The coherence condition (which requires a low $\leq$ average $\leq$ high pattern of estimates) was violated by three modelers. Their estimates were not provided to recipients, and their data are excluded from all analyses below.

Before providing estimates of YT experience, modelers provided WTP judgments for one or two job candidates. Mean valuations are depicted in Figure 3A. Separate evaluations of A exceeded separate evaluations of B (47.3 vs. 43.0 thousand), unpaired $t(75) = 2.43$, $p = .02$. In contrast, joint evaluations of A fell well below joint evaluations of B (40.6 vs. 50.1 thousand), paired $t(37) = 6.32$, $p < .0001$. Modelers in Experiment 1 exhibited a typical JSR.

What inferences did modelers draw from the option-samples they received? Figure 4 depicts mean values for the low, average, and high YT estimates of modelers who saw each evaluation set. When the evaluation set was {A}, YT experience was judged to range from 9.6 to 23.3, with an average of 15.2 YT programs. When the evaluation set was {A, B}, YT experience was judged to range from 20.0 to 68.2, with an average of 39.9 YT programs. Finally, when the evaluation set was {B}, YT experience was judged to range from 57.2 to 130.1, with an average of 85.1 YT programs. When low, average, and high YT estimates are separately analyzed, all pairwise comparisons between evaluation-set conditions are significant ($ps < 0.001$). Modelers' beliefs about
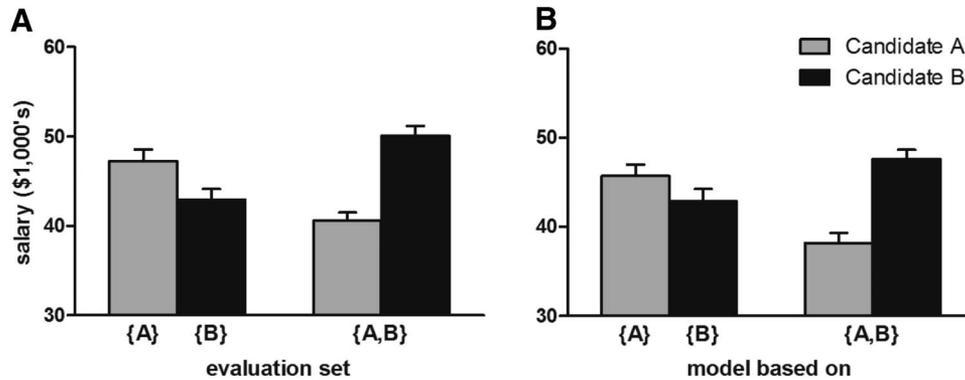
Figure 3. Mean willingness-to-pay judgments of modeler participants (A) and yoked recipient participants (B) in Experiment 1. Standard error bars are shown.

the distribution of YT experience exhibited a strong assimilation effect, shifting markedly toward the statistics of the option-samples they received.

How did the inferences of modelers impact the evaluations of recipients? Figure 3B depicts the mean WTP judgments of recipients as a function of the source of the YT background information they received and the single option they evaluated. A familiar pattern of evaluations emerges: Evaluations of A based on a YT model inferred from sampling {A} exceeded evaluations of B based on a model inferred from sampling {B}, though the difference fell short of significance (45.7 vs. 42.9 thousand), $t(71) = 1.53$, $p = .13$. In contrast, evaluations of A based on a model inferred from sampling {A, B} fell well below evaluations of B based on a model inferred from sampling {A, B} (38.2 vs. 47.6), $t(74) = 6.05$, $p < .0001$. The inferences of modelers led to the familiar JSR pattern in the valuations of recipients.

According to the options-as-information model, this JSR pattern is caused by a contrast effect, in which poorer posterior models lead to higher valuations of options. Confirming this prediction,
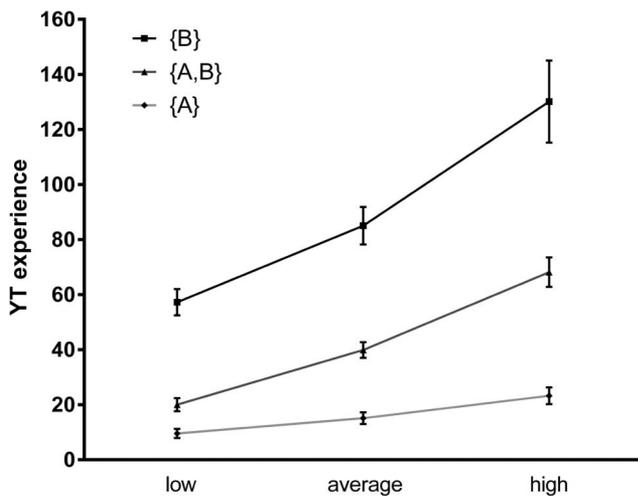
Figure 4. Mean estimates for the low, average, and high YT values of typical job candidates, provided by modeler participants in Experiment 1. Standard error bars are shown.

recipients' valuations of A were higher when based on the inferior (i.e., downward-shifted) models inferred from {A} rather than on models inferred from {A, B}, $t(72) = 4.43$, $p < .0001$. Similarly, WTPs for B were lower when based on the superior (i.e., upward-shifted) models inferred from {B} rather than on models inferred from {A, B}, $t(73) = 2.74$, $p < .01$. Improved models of the market lead to diminished valuations of individual options.

Discussion

The results of Experiment 1 corroborate the core predictions of the options-as-information model of JSRs. First, different samples led to very different inferences. Modelers exhibited a strong assimilation effect, with YT estimates gravitating toward the statistics of the option-sample they encountered (Figure 4).[2] Second, recipients' evaluations were driven by contrast to the background information they received, with better models leading to lower valuations of options. Compounding the effect of sample on model, and the effect of model on WTP, we obtain a valuation pattern (Figure 3B) that qualitatively reproduces the JSR in the joint-separate design (Figure 3A).

These results have both normative and psychological implications. In normative terms, participants' joint and separate evaluations pass the test of coherence that is appropriate when prior knowledge of the option space is limited. In psychological terms, the fact that inferences from evaluation sets have the same effect as the evaluation sets themselves suggests that the JSR may be parsimoniously explained as a product of sampled-based inference. Note that all recipients evaluated a single option on the basis of equally detailed background information. The findings suggest that beliefs about the market, rather than evaluation mode or the mere quantity of information, are the critical factor in determining the evaluation of options.

---

[2] A possible concern about the YT estimates in Experiment 1 is that they were provided only after participants had expressed their WTP judgments, raising the possibility of cross-response contamination. However, in an unpublished study including a similar modeler component but with no WTP judgment preceding it, we found a pattern of YT estimates almost identical to that depicted in Figure 4.

Evaluation Reversal Without Preference Reversal

JSRs are classed as preference reversals, in which the preference ranking of d and e flips between joint and separate evaluation. (Recall that d is the option that dominates on the difficult-to-evaluate attribute, while e excels on the easy-to-evaluate attribute.) We argue below that this classification is mistaken: JSRs are not preference reversals.

JSRs are usually conceived as "effects of evaluation mode" (joint vs. separate). However, note that evaluation set ($\{d\}$, $\{e\}$, or $\{d, e\}$) rather than mode is the independent variable in these experiments. In the normative model (Figures 1 and 2), each evaluation set gives rise to a different posterior model, and hence to a distinct preference order. Consequently, we cannot assume that preferences are identical in the two separate evaluation conditions ($\{d\}$ and $\{e\}$).

The panels of Figure 5 illustrate how the same JSR can result from very different underlying patterns of preference. In the figure, $\geq_{\{d\}}$, $\geq_{\{e\}}$, and $\geq_{\{d,e\}}$ denote the underlying preference orders of a hypothetical DM in the three conditions of a joint-separate experiment. The orders are represented by vertical arrays, with higher points preferred to lower points. Figure 5A depicts the basic observational facts: $WTP(e|\{e\}) > WTP(d|\{d\})$ and $WTP(d|\{d, e\}) > WTP(e|\{d, e\})$. Figures 5B and 5C depict two conflicting interpretations of these facts.

If we treat the separate evaluation conditions as interchangeable, then, as Figure 5B illustrates, we can translate monetary indifference points between $\geq_{\{d\}}$ and $\geq_{\{e\}}$, leading to the conclusion that "e is preferred to d in separate evaluation." However, the data in a joint-separate experiment are equally compatible with the underlying preference orders in Figure 5C, in which d is definitely preferred to e in every condition, and hence there is no preference reversal over the focal options. In fact, this underlying pattern of preferences results from an options-as-information analysis, as we now show.

Consider an arbitrary set V of options, and a specific option x, not necessarily in V. Now obtain a new set V' by lowering (i.e., degrading) the attribute values in V. By normative assimilation, the rational DM will infer a poorer distribution of the difficult-to-evaluate attribute when V' rather than V is sampled. As the posterior distribution shifts lower, the DM's valuation of x will rise by normative contrast. That is, the receipt of an inferior sample leads to a "rising tide of valuation" that lifts all options. For example, a candidate who has programmed 30 YT programs emerges as better relative to the model inferred from sampling $\{A\}$ (10 programs) rather than $\{B\}$ (70 programs).

When the evaluation set in the joint-separate design is $\{d, e\}$, we observe that d is valued above e. Sampling $\{e\}$ leads to a poorer posterior model and should therefore increase the DM's private valuations of both d and e. However, there is no reason to think that the ordering of d and e will reverse. Sampling $\{d\}$ leads to a superior posterior model, and thus should lower the valuations of d and e, again without reversing their order. In the joint-separate design, we thus elicit a separate evaluation of e ($WTP[e|\{e\}]$) when d and e are both at "high tide," and we obtain a separate evaluation of d ($WTP[d|\{d\}]$) when d and e are both at "low tide." Hence e receives a higher separate evaluation than d. This creates the illusion that "e is preferred to d in separate evaluation" when in fact the DM definitely prefers d to e in all conditions (Figure 5C).
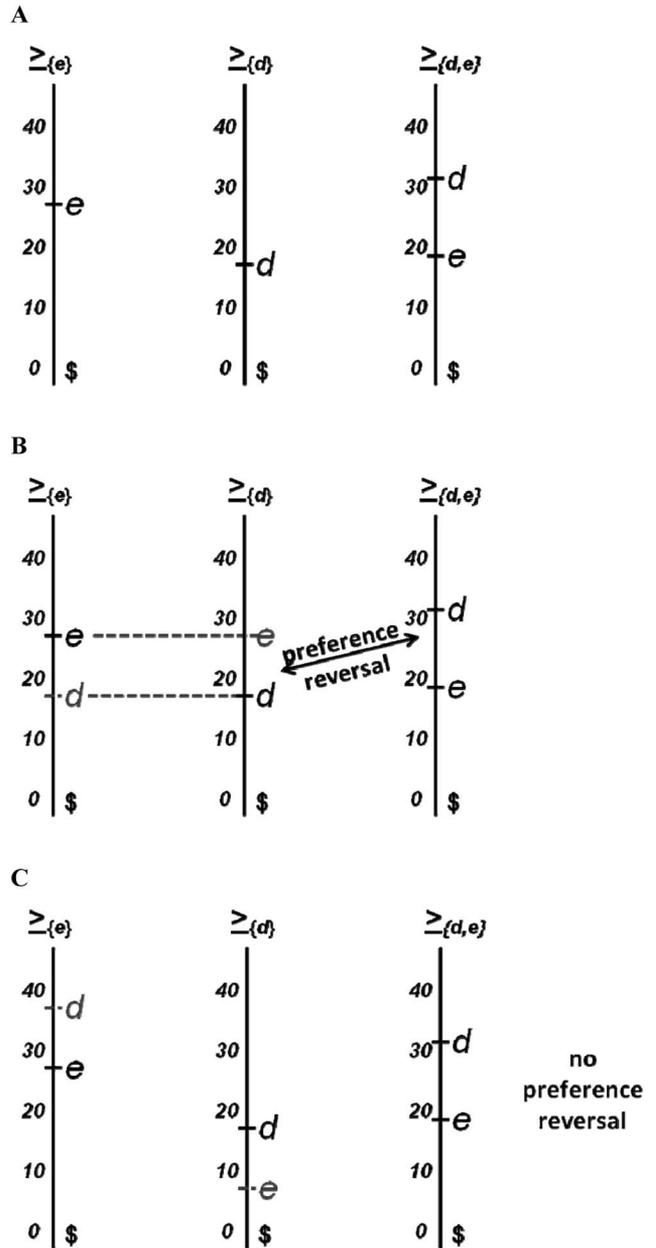
Figure 5. Actual data (A) and possible interpretations (B and C) for a prototypical joint-separate experiment. The preference orders in the three experimental conditions are represented by vertical lines, in which higher points are preferred to lower points. In reconstructing preference orders from judgments, we assume that preferences are transitive, more money is preferred to less, and strict indifference obtains between an option and its stated willingness-to-pay. The pattern of preference orders in B follows if the two separate evaluation conditions are treated as interchangeable. The nonreversing pattern of preference orders in C emerges in the normative model of inference and evaluation.

For example, in the programmer problem, sampling $\{A\}$ rather than $\{B\}$ suggests a downward-shifted posterior distribution relative to which both candidates are better. Thus, the rational DM's private valuations of *both* A and B are higher when $\{A\}$ rather than

{B} is sampled. This explains why separate evaluations are higher for A than B, but does not imply that A is ever preferred to B.

The results of Experiment 1 suggest that the JSR is parsimoniously explained by the inferences people draw from evaluation sets, and the valuations they provide in light of these inferences. Accordingly, to determine the implicit ordering of d and e in (for example) $\geq_{\{d\}}$, we must go beyond the joint-separate design, and ask how e is evaluated when participants' beliefs about the market correspond to what would be inferred from sampling $\{d\}$. We ask this question in Experiment 2. The answer suggests a new conception of JSRs, as evaluation reversals without preference reversal.[3]

The foregoing analysis reflects the fact that *preference*, a within-subjects relation, is not captured in between-subjects comparisons of evaluations. It echoes Birnbaum's (1999) compelling warning about the pitfalls of between-subjects comparison. In Birnbaum's study, the number 9 received higher ratings of subjective magnitude than 221 when the numbers were judged in isolation. Presumably, the numbers lead to different interpretations of the ambiguous magnitude scale—9 suggests an interpretation of the scale in which 10 counts as "large," while 221 suggests an interpretation in which 10 counts as "small." Similarly, different separate evaluation contexts lead to different interpretations of the distribution of the difficult-to-evaluate attribute, and thus to different assignments of value to options. In both cases (ambiguous response scale or poorly known attribute), the comparison of judgments across different interpretations is not equivalent to a comparative judgment within any interpretation.

Experiment 2

In Experiment 2, the posterior YT estimates of modelers from Experiment 1 were provided as background information to a new group of participants. Each Experiment 2 participant evaluated one option, based on background information about the YT distribution drawn from an Experiment 1 participant who saw one option. Experiment 2 has four conditions, corresponding to the four possible combinations of model ({A}-based or {B}-based) and evaluated option (A or B). In this way, we unconfound the target of evaluation from the base of inference. We can thus learn the valuation of an option relative to the posterior model inferred when the other option is sampled alone.

Method

Two hundred twenty-eight UCSD undergraduates participated in Experiment 2 for partial credit in psychology courses. Data were not obtained for one participant due to a computer error.

The background blurb and evaluation stimuli (programming candidates) were identical to those in Experiment 1. All participants in Experiment 2 were "recipients." Each participant received background information about the YT distribution which, unbeknownst to the participant, was derived from an Experiment 1 modeler participant. In addition, each participant saw and evaluated a single option—A or B. There were four conditions in Experiment 2: *A|{A}-recipients* ($n = 58$) evaluated A and received background information derived from a modeler who saw {A}. Other conditions tested *B|{A}-recipients* ($n = 55$), *A|{B}-recipients* ($n = 58$), and *B|{B}-recipients* ($n = 56$).

The A|{A} and B|{B} conditions directly replicate the corresponding recipient conditions from Experiment 1. The B|{A} and A|{B} conditions are new and allow us to ask about participants' evaluation of each option, when their information about the YT distribution corresponds to what would be inferred from evaluating the other option. Participants were instructed to type a number between 30 and 60 to indicate how many thousands of dollars they would be willing to pay the candidate in the table.

We used the following procedure for sampling Experiment 1 YT estimates. For each A|{A}-recipient in Experiment 2, the YT estimate of a single {A}-modeler from Experiment 1 was selected at random and provided as background information to this new participant. The next B|{A} recipient on the same computer received the same background information. This ensured that, while {A}-based estimates were randomly sampled from Experiment 1 modelers, the distribution of received estimates would be the same for the A|{A} and B|{A} conditions. The same procedure was used for the A|{B} and B|{B} conditions. Due to a programming error, three Experiment 1 modelers were inadvertently double-counted in creating the parent set from which {A}-based estimates were randomly sampled in Experiment 2 and also in creating the parent set of {B}-based estimates. However, the discrepancy between the Experiment 1 estimate populations and the sets from which estimates were drawn in Experiment 2 was negligible.

Results

Four WTP responses were range-adjusted in Experiment 2. (See Experiment 1 Method for details.)

Mean valuations for each option under each posterior model are depicted in Figure 6. When posterior models were inferred from sampling {A}, mean WTP was higher for B than for A (48.6 vs. 44.9 thousand), $t(111) = 2.40$, $p = .02$. Likewise, when posterior models were inferred from sampling {B}, mean WTP was higher for B than A (41.0 vs. 36.9 thousand), $t(112) = 3.05$, $p < .01$. The ordering of the focal options is the same for both model types, and corresponds to the ordering (B > A) seen under joint evaluation in Experiment 1.

The inset in Figure 6 shows how this pattern of valuations creates the illusion that "A is preferred to B in separate evaluation." The inset reproduces the subset of the data in which an option was evaluated on the basis of a model inferred from sampling that option. According to the options-as-information model, this is the process that leads to separate evaluations in the standard design. Mean WTP for A based on a model inferred from {A} (44.9 thousand) exceeded mean WTP for B based on a model inferred from {B} (41.0 thousand). This pattern reproduces that

---

[3] There is a technical sense in which the standard JSR might be said to involve a preference reversal, but not one involving the ordering of the focal options (i.e., A and B in Experiments 1 and 2). The same may be said of any experiment in which a between-subjects variable influences valuations of a fixed option a. In particular, suppose the valuation $WTP_1(a)$ in Condition 1 exceeds the valuation $WTP_2(a)$ in Condition 2, and let m be any monetary amount between $WTP_1(a)$ and $WTP_2(a)$. If the DM is strictly indifferent between an option and its stated cash equivalent, it follows that a is preferred to m in Condition 1 while m is preferred to a in Condition 2. Of course, JSRs are generally classified as preference reversals because they are believed (erroneously, we argue) to imply a reordering of the focal options, and not because they demonstrate reversals between each option and the monetary amounts in between its two stated WTPs.
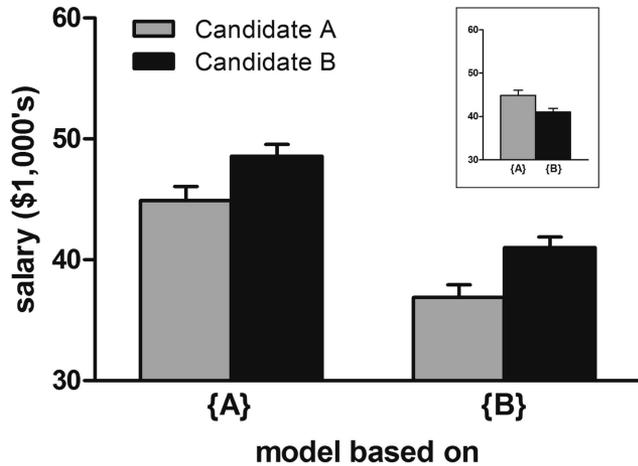
Figure 6. Mean willingness-to-pay judgments in Experiment 2. Regardless of whether the model they receive is inferred from {A} or {B}, recipients' valuations are higher for B than A. However, as the inset shows, the valuation of A based on a model inferred from {A} exceeds the valuation of B based on a model inferred from {B}.

found in the corresponding conditions of Experiment 1, and in Experiment 2 this difference was significant, $t(112) = 2.67$, $p < .01$. Experiment 2 thus reproduces the "separate evaluation" advantage for A over B,[4] while also demonstrating that B is valued above A in every information state.

Discussion

Experiment 2 unconfounded the base of inference from the target of evaluation. We found that sampling {A} leads to a rising tide of evaluation that lifts both A and B by similar amounts. This explains why A receives higher separate evaluations than B in the joint-separate design. However, the ordering of separate evaluations does not correspond to any preference order, as B receives a higher WTP relative to every posterior model.

Taken together, the results of Experiments 1 and 2 can be summarized as follows. First, the sampling of options has predictable effects on beliefs about the market—the market appears worse when {A} rather than {B} is sampled. Second, beliefs about the market have predictable effects on the valuation of options—both A and B are valued more highly when the market appears worse. Putting these two effects together, we obtain a standard JSR (Experiment 1). However, we obtain no reversal of preference between A and B (Experiment 2).

Varieties of Joint Evaluation

In the standard joint-separate design, two options are seen in the joint evaluation condition. However, as we noted earlier, the evaluability hypothesis makes predictions for the more general design, in which participants may see {d}, {e}, or a set U with {d, e} ⊆ U. In this section, we extend the options-as-information analysis to designs with larger joint-evaluation sets. We show that the normative model predicts (a) systematic violations of the EI and (b) a novel class of rational preference reversals.

The Evaluability Inequality Revisited

Recall that the evaluability hypothesis predicts the EI, which states that the joint-evaluation advantage for d over e exceeds its separate-evaluation advantage. Formally, $\Delta_{joint} = WTP(d|U) - WTP(e|U) > WTP(d|\{d\}) - WTP(e|\{e\}) = \Delta_{sep}$.

The normative model also predicts the EI in the standard two-option design. This is because model-normalized levels of the difficult-to-evaluate attribute hardly differ when a single option is sampled (leading to lower Δ_{sep})—but, relative to the common posterior model inferred when {d, e} is sampled, d is superior to e (leading to higher Δ_{joint}). However, matters are different in multi-option designs with $U \neq \{d, e\}$. In this case, the additional elements of U provide additional information about the variability of the difficult-to-evaluate attribute. For example, consider two possible sets of sampled values on the difficult-to-evaluate attribute: $D_1 = \{10, 20, 12, 17\}$ and $D_2 = \{10, 20, 240, 350\}$. These two sets suggest different posterior models of market variability for the attribute. The fixed numerical difference of 10 between the first two values translates into a large percentile difference in the model inferred from D_1, and to a small percentile difference in the model inferred from D_2. In general, if the difference between d and e on the attribute is very small relative to the variability in U, then the model-relative standing of d will be drawn toward the model-relative standing of e, and Δ_{joint}, like Δ_{sep}, will diminish with increasing assimilation. The EI will no longer be stable in this case. Alternatively, if the difference between d and e is very large relative to the variability in U, increasing assimilation will increase Δ_{joint}, inflating the EI.

The options-as-information model thus identifies boundary conditions for the evaluability hypothesis: In multi-option designs, the EI will hold up when the values of the focal options on the difficult-to-evaluate attribute are distant relative to the variability in U, but the EI will become unstable when the focal options are close relative to the variability in U. Experiment 3 below provides a test of this prediction.

Rational Preference Reversals

When the distance between focal options is small relative to the variability of U, the normative model makes a further prediction. Consider the CD changers in Table 3, each defined by the number of CDs carried and by a measure of sound distortion (the "Sound Distortion Index" or "SDI"). Lower values of SDI indicate higher sound quality. CD changers A and B are identical to the stimuli in Study 4 of Hsee (1996), except that we have replaced the authentic but unfamiliar measure of distortion used there with the invented measure of SDI. In that study, some participants saw {A}, some saw {B}, and the remaining participants saw $U = \{A, B\}$. A typical JSR was found. A received higher separate evaluations than B, but participants paid more for B in joint evaluation.

---

[4] To confirm the reliability of the effect across both experiments, we ran a 2 × 2 ANOVA with experiment (1 or 2) and recipient-condition (A|{A} and B|{B}) as between-subjects factors, and stated WTP as the dependent variable. The analysis reveals a main effect of condition, with higher WTPs provided by A|{A} than B|{B} recipients, $F(1, 183) = 8.19$, $p = .005$. Neither the main effect of experiment nor the experiment-by-condition interaction is significant ($ps > 0.25$).

Table 3
Stimuli for Experiment 3

	CD capacity	Sound quality
CD changer A	20	SDI = .01%
CD changer B	5	SDI = .003%
CD changer C	15	SDI = 39.42%
CD changer D	10	SDI = 58.77%

Note. SDI = Sound Distortion Index.

Now consider the expanded evaluation set $U' = \{A, B, C, D\}$. According to the normative model, sampling U' suggests a posterior SDI distribution relative to which the numerical difference between A and B is negligible. With increasing assimilation, then, the valuations of A and B relative to the posterior model inferred from sampling U' are differentiated only by CD capacity, which favors CD changer A. Thus A should receive a higher valuation than B when U' is sampled. In this case, the option that is inferior on the difficult-to-evaluate attribute should be *preferred* in joint evaluation.

Consider a design that compares a "double evaluation" condition, in which $\{A, B\}$ is evaluated, with a "quadruple evaluation" condition in which $\{A, B, C, D\}$ is evaluated. The options-as-information model predicts a rational preference reversal in this design, with B valued above A in double evaluation while A is valued above B in quadruple evaluation. This evaluation reversal would be a true preference reversal, unlike the JSR in the standard two-option design. Because each participant evaluates A and B at the same time, the order of preference can be inferred from the order of valuation.

The predicted effect could also be explained by range-frequency theory (Parducci, 1965, 1995). In a range-frequency model, the subjective value assigned to an objective attribute is a weighted average of (a) its relative rank in the stimulus set, and (b) its relative position in the range defined by the extreme stimuli in the set (cf. Wedell & Pettibone, 1996). Relative to the $\{A, B\}$ condition, A rises in proportional rank and range-relative position for SDI in the $\{A, B, C, D\}$ condition, but B does not. Thus, a range-frequency analysis, like the normative model, would predict higher evaluations for A (but not B) in quadruple relative to double evaluation.

Note, however, that the normative model makes specific predictions about the relationship between valuations across a range of contexts. For example, the normative model predicts that $WTP(B|\{A,B,C,D\}) > WTP(B|\{A,B\}) > WTP(B|\{B\})$, as the inferred SDI distribution progressively improves from the first context to the last. However, the range-relative position and proportional rank of B are maximal in both the $\{A, B, C, D\}$ and $\{A, B\}$ contexts and undefined in the $\{B\}$ context (where there is no range). To reproduce these normative predictions, a range-frequency account would have to assume that invoked subjective contexts differ from the presented objective contexts. Moreover, the implicit contexts suggested by the presented contexts would presumably have to be similar to the posterior models inferred from the evaluation sets. While the notion that stimulus sets evoke richer implicit contexts is consistent with range-frequency theory (Parducci, 1995), we think the normative model provides a more direct and parsimonious account of the above inequalities.

The predicted preference reversal is related to several known effects of "global context"—i.e., of attribute values seen on earlier experimental trials, or in an extended set of background options. When, in the global context, one attribute range is contracted and the other is expanded, preferences shift toward the option that excels on the contracted attribute (Mellers & Cooke, 1994, 1996; Stewart, 2009). Such effects may weaken when attributes are better-known (Mellers & Cooke, 1994). Similarly, when only one attribute is difficult to evaluate, simultaneous contraction of both ranges shifts preferences toward the option that is best on this attribute (Yeung & Soman, 2005). Prior accounts have attributed these findings to range-frequency theory or to Stewart, Chater, and Brown's (2006) decision by sampling theory. We note that these effects can also be explained by the options-as-information model: In posterior models inferred from a contracted attribute range, a fixed objective difference translates into a larger model-normalized difference, and this effect should be most pronounced when attributes are least familiar.

In summary, the options-as-information model explains why the evaluability hypothesis makes correct predictions in two-option designs. Both accounts predict the EI ($\Delta_{joint} > \Delta_{sep}$) in this case. However, unlike the evaluability hypothesis, the normative model entails a breakdown of the EI when the variability of the joint-evaluation set is large relative to the numerical spacing of the focal options. Going further, the model predicts a novel (and rational) preference reversal, in which the ranking of focal options differs in different varieties of joint evaluation. These predictions are tested in Experiment 3.

Experiment 3

Method

Two hundred twenty-five UCSD undergraduates participated in Experiment 3 for partial credit in psychology courses. One participant who responded with a WTP range when a single number was requested was excluded. A paper-and-pencil questionnaire asked participants to:

> Imagine that you own a large CD collection, and you decide to buy a new CD changer at an electronics store. (A CD changer is a CD player that holds multiple CDs at the same time.) You plan to spend between $150 and $300 on a CD changer.

A table displayed specifications for between one and four CD changers. Each product was defined in terms of "CD capacity" and "Sound quality," the latter indexed by a measure called "SDI." Participants were told that CD capacity refers to the number of CDs the changer could hold. They were also informed that "SDI" stands for "Sound Distortion Index," and that "The smaller the SDI value, the better the sound quality."

The CD changers in Experiment 3 were drawn from the set $\{A, B, C, D\}$ defined in Table 3. Participants in a *quadruple evaluation* condition ($n = 57$) saw all four CD changers. Participants in a *double evaluation* condition ($n = 56$) saw only CD changers A and B. Participants in a *{A} evaluation* condition ($n = 55$) and in a *{B} evaluation* condition ($n = 56$) saw only one product. The letter name of the CD changer was omitted in the two separate evaluation conditions. Below the stimulus table, between one and four WTP questions were printed. Participants were instructed that they could answer with any value between 150 and 300.

Results

Eight individual WTP responses for the focal options (CD changers A and B) were range-adjusted, as were nine responses for the nonfocal options (C and D). (See Experiment 1 Method for details.)

Figure 7 depicts mean WTP for the focal options in the four evaluation-set conditions. In the separate-evaluation conditions, WTP for A when {A} was presented exceeded WTP for B when {B} was presented (205.5 vs. 187.7), $t(109) = 2.27$, $p = .03$. In contrast, in the double-evaluation condition, WTP was lower for A than for B (193.1 vs. 205.1), $t(55) = 1.49$, $p = .14$. These comparisons replicate Hsee (1996), although the joint-evaluation advantage for B fell short of significance. In the quadruple-evaluation condition, WTP was once again higher for A than B (261.6 vs. 218.3), $t(56) = 5.25$, $p < .0001$, as predicted by the options-as-information model. Not surprisingly, WTPs were very low for the further options C (174.0) and D (151.9) in this condition.

The options-as-information model predicts a preference reversal, in which A is valued above B more often in the {A, B, C, D} condition than in the {A, B} condition. The analyses reported above focused on mean WTP. In comparing two joint evaluation conditions, a more direct test of preference reversal determines rates of preference by considering how often individual participants value one option over another. By this measure, A was preferred to B 30.4% of the time in the {A, B} condition, while B was preferred to A 57.1% of the time (with 12.5% of participants assigning equal value to the two options). In the {A, B, C, D} condition, A was preferred to B 77.2% of the time and B was preferred to A 15.8% of the time (with an indifference rate of 7.0%). A's preference share was thus significantly above 50% when {A, B, C, D} was sampled ($p < .0001$) and significantly below 50% when {A, B} was seen ($p < .01$). Participants in Experiment 3 exhibited the preference reversal predicted by the normative model.

The evaluability hypothesis predicts that the EI should hold in the comparison of single and quadruple evaluation. That is,

Figure 7. Mean willingness-to-pay judgments for the two focal CD changers across the four evaluation sets in Experiment 3. Standard error bars are shown.

$\Delta_{joint} = WTP(B|U) - WTP(A|U)$ should be greater than $\Delta_{sep} = WTP(B|\{B\}) - WTP(A|\{A\})$ when $U = \{A, B, C, D\}$. Contrary to the evaluability hypothesis, we found that $\Delta_{joint} = -43.2$ was less than $\Delta_{sep} = -17.7$ for the quadruple evaluation set. To confirm the significance of this EI violation, we conducted a 2 (option: A or B) $\times$ 2 (set size: single or quadruple) analysis of variance (ANOVA), in which the dependent variable was stated WTP.[5] The ANOVA revealed main effects of both option, with A receiving higher valuations than B overall, $F(1, 221) = 27.66$, $p < .001$, and set size, with both A and B receiving higher valuations for quadruple relative to single sets, $F(1, 221) = 55.99$, $p < .001$. Importantly, the interaction between option and set size was significant, $F(1, 221) = 4.86$, $p = .03$. That is, B's WTP advantage over A in "quadruple evaluation" exceeded its advantage across the separate evaluation conditions.

Discussion

The evaluability hypothesis and the options-as-information model make different predictions in Experiment 3. The former predicts that the EI will hold in the comparison of single and quadruple evaluation. Instead, and consistent with the normative model, we found a reversal of the EI ($\Delta_{sep} > \Delta_{joint}$) when the joint evaluation set was {A, B, C, D}. Furthermore, the normative model predicts a rational preference reversal across joint evaluation conditions, with B preferred to A in double evaluation while A is preferred to B in quadruple evaluation. The predicted preference reversal was confirmed. In this analysis, within-subjects WTP orderings in a fixed information state provide a valid measure of individual preference. The reversal emerges at the group level, when we compare, between-subjects, rates of individual preference across the joint evaluation conditions.

We noted earlier that the options-as-information model predicts a finer-grained pattern of evaluations in Experiment 3. Relative to the {A} condition, valuations of A should be lower when {A, B} is sampled (upward-shifted posterior model) but higher when {A, B, C, D} is sampled (downward-shifted model). Relative to the {B} condition, valuations of B should be higher in the {A, B} condition (downward-shifted model) and higher still in the {A, B, C, D} condition (further downward-shifted model). This exact pattern of evaluations was seen in Experiment 3.[6] While these findings are not incompatible with a range-frequency account (Parducci, 1995), they are specifically predicted—and, we think, most parsimoniously explained—by the normative model.

It is instructive to fully unpack the findings of Experiment 3 from the options-as-information perspective. Although separate evaluations are higher for A than B, the model implies that B is preferred to A in both separate evaluation conditions. In terms of our earlier notation, where $\geq_{\{\cdots\}}$ denotes the DM's underlying preferences when set {$\cdots$} is sampled, we have $B >_{\{A\}} A$ and $B >_{\{B\}} A$. In double evaluation, the same relation holds: $B >_{\{A,B\}} A$. But in quadruple evaluation, DMs infer that the SDI difference
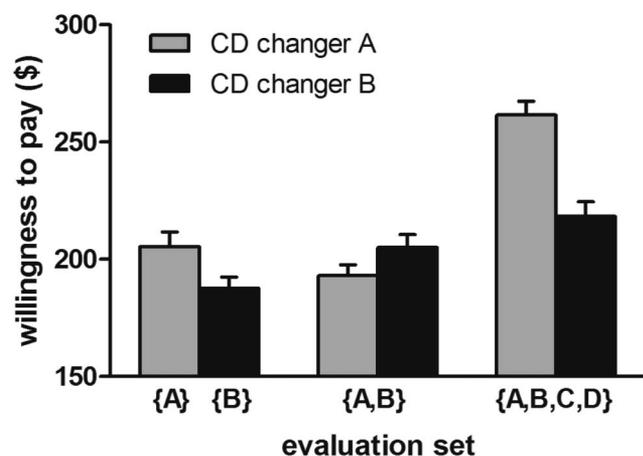
---

[5] Note that this analysis does not have a repeated-measures component—i.e., it treats evaluations of A and B as two independent samples in both the single and quadruple contexts.

[6] All pairwise comparisons were in the direction predicted by the normative model. All were individually significant, with the exception of the A|{A} *v.* A|{A, B} and B|{A, B} *v.* B|{A, B, C, D} comparisons ($ps < 0.12$).

between A and B is negligible relative to the empirical variability of the attribute. The result is a preference reversal, relative to all other conditions, with $A >_{\{A,B,C,D\}} B$.

Thus, in addition to the demonstrated preference reversal between the {A, B} and {A, B, C, D} conditions, the options-as-information model implies a theoretical preference reversal between the {A, B, C, D} condition and each separate evaluation condition ({A} and {B}). That is, the option that is superior on the difficult-to-evaluate attribute is preferred in each *separate* evaluation condition, while the option that is superior on the easy-to-evaluate attribute is preferred in the *joint* evaluation condition. These preference reversals invert the traditional picture of separate and joint evaluation.

Application to Related Paradigms

The evaluability hypothesis has been deployed to explain findings in several related experimental paradigms. In this section, we survey some of these findings—concerning the effects of reference information (Hsee & Leclerc, 1998; Willemsen & Keren, 2004), the "less-is-better" effect (Hsee, 1998; List, 2002), and metacognitive judgments of attribute impact (Hsee, 2000). We suggest that these phenomena are naturally explained by the options-as-information model.

Effects of Reference

Hsee and Leclerc (1998) studied the *combined* valuation of two options, in choice problems where both attributes could be difficult-to-evaluate. The question was whether the sum of separate evaluations (ΣSE) would be greater or less than the sum of joint evaluations (ΣJE). The studies examined the impact of "reference information" on the relation between sums. The "reference" was operationalized either as the typical participant's prior beliefs about the average values of the attributes or as a description of another product recently bought by a friend. When the reference was inferior to both options, the researchers found that ΣSE > ΣJE. However, when the reference was superior to both options, ΣJE > ΣSE This finding was replicated by Willemsen and Keren (2004).

The normative model predicts this pattern of results. Note that, whether it is an uncertain prior belief or an additional option sampled, "the reference" should make a larger contribution to the DM's posterior model when only one option is presented than when two options are sampled. Therefore, when both options lie on the same side of the reference, the posterior distribution will be closer to the reference, on average, for one-option than two-option samples.

When the reference is inferior, the model inferred from one-option samples will thus be poorer than the model inferred from two-option samples. By normative contrast, overall valuations will be higher for the one-option samples: ΣSE > ΣJE. Conversely, for a superior reference, the model inferred from the one-option samples will be better than that inferred from the two-option sample; therefore, by normative contrast, ΣJE > ΣSE. Reference effects naturally fall out of the normative model, because larger samples more effectively dilute reference information.

The "Less-Is-Better" Effect

Hsee (1998) reported experiments in which an option receives a higher separate evaluation than a second option which apparently dominates it. A representative study involved two dinnerware sets in a clearance sale—one with 24 pieces, and a second with 40 pieces, of which nine were broken (i.e., a total of 31 intact pieces). The smaller set received lower joint evaluations but higher separate evaluations. A similar effect was found in a market study by List (2002). Reflecting on the order of separate evaluations in these studies, Kahneman (2011) wrote that "from the perspective of economic theory, this result is troubling," because "adding a positively valued item to the set can only increase its value." (p. 161).

However, an options-as-information analysis explains the "less-is-better" effect while defusing the apparent violation of dominance. In the dinnerware study, the total size of the set(s) sampled may support an inference about the distribution of set sizes on the market. The number of *intact* pieces in a set, relative to this inferred distribution, determines its valuation. When sets are sampled alone, the larger set has a lower model-normalized number of intact pieces, relative to the upward-shifted distribution inferred from its total size (40 pieces). The larger set thus receives a lower separate evaluation. However, relative to every fixed model, more intact pieces are preferred to fewer.

The above discussion ignores further inferences that may also contribute to effects in this paradigm. For example, one may naturally suppose that a clearance sale featuring a partly broken dinnerware set is likely to involve lower quality items in general. This would depress separate evaluations of the larger but not the smaller set. In joint evaluation, the two sets are part of the same sale, and inferences about general quality apply to both sets equally. Similar inferences may have played a role in List's (2002) study of buyers and dealers in a baseball card market. In this study, a base set of 10 mint-condition cards received lower separate evaluations than a larger set in which three low-quality cards were added. This effect may be explained by inferences from options, on the plausible assumptions that (a) evaluators are uncertain about the value of individual cards, and (b) card value is positively correlated within sets sold on the market. This analysis predicts that effect sizes should diminish with increasing prior knowledge of card value. Indeed, dealers, who were presumably more knowledgeable than buyers, exhibited markedly diminished effects in List's (2002) study.

Metacognitive Judgment

In a study reported by Hsee (2000), candidates for a programming position varied in personal appearance (unkempt or well-groomed) and experience (number of programs written). Each participant evaluated a single candidate, with levels of appearance and experience independently varied in a between-subjects design. Participants also provided metacognitive reports concerning the attributes' relative influence on their judgments. The findings were that (a) level of appearance influenced judgments, (b) level of experience did not influence judgments, but (c) a substantial majority of participants reported that they were more influenced by experience than appearance. This was deemed a metacognitive error, as "beliefs were grossly inconsistent with the actual influence of the attributes" (Hsee, 2000, p. 561).

However, if *model-normalized* attribute values play a causal role in the evaluation process, then participants' metacognitive beliefs are consistent with their valuations. Extrapolating from Experiment 1, each candidate's experience level in this experiment would be normalized to the same value (about average) relative to the posterior model of experience inferred when that candidate is seen alone. As a result, normalized experience does not vary in this experiment, and its influence on valuation cannot be determined. Nonetheless, metacognitive reports (that experience was weighted more heavily than appearance in the valuation process) align with the options-as-information model, and the model predicts the pattern of judgments that was observed.

General Discussion

When prior knowledge of attributes is limited, options are informative. Rational choice then emerges from a two-stage process of inference and evaluation. We showed that, when a single important attribute is poorly known, this process can lead to JSRs in multi-attribute evaluation. Three experiments supported the descriptive validity of this normative analysis. Inferences drawn from evaluation sets have the same effect as the evaluation sets themselves (Experiment 1). That is, sample-based inferences suffice to reproduce the pattern of valuations in a JSR, even when evaluation mode is always separate and the total quantity of information is fixed. However, JSRs are not preference reversals; the ordering of evaluations is the same relative to all inferred models (Experiment 2). Furthermore, the model explains the break-down of the EI and predicts the emergence of novel preference reversals, in expanded designs involving more than two options (Experiment 3). The model also provides a natural explanation of behavior in several variants of the joint-separate paradigm. It can account for effects of reference and the "less-is-better" effect, and it suggests that metacognitive reports about the impact of attributes may be faithful descriptions of the process of evaluation.

The options-as-information model explains why the evaluability hypothesis works in standard designs and identifies its boundary conditions in multi-option experiments. It also suggests a revised view of the decision theory and rationality of these effects. Hsee and Zhang (2010, p. 348) expressed a commonly held assessment of JSRs when they wrote, "Intuitively, two items (e.g., two job candidates) evaluated comparatively (i.e., in JE) or separately (i.e., in SE) should yield the same preference ranking. If Candidate A is rated more favorably than Candidate B in JE, then A should also be rated more favorably than B in SE." In reality, preference ranking often reverses between JE and SE." The present analysis supports a different assessment, according to which some widely studied JSRs are neither preference reversals nor counternormative.

The evaluability hypothesis states that the impact of difficult-to-evaluate attributes is greater in joint than in separate evaluation. In some expositions, this "impact" hypothesis is translated into a statement about attribute weights; that is, it is assumed that joint and separate evaluation processes make use of distinct underlying attribute weights (e.g., Bazerman et al., 1999; Shafir & LeBeouf, 2004). The present analysis makes clear that JSRs do not imply inconsistent weighting of attributes. In the rWADD model, fixed weights are assigned to model-normalized attributes, and simple

belief updating is sufficient to explain the emergence of a JSR. More generally, JSRs do not imply that joint and separate evaluation tap into different cognitive processes.

Recently, Hsee and Zhang (2010) proposed GET, which refines and extends the evaluability hypothesis. Its core postulate is a general positive relationship between attribute evaluability and value sensitivity, where the latter is defined as the difference in valuation that results from a fixed objective increment on the attribute. The theory retains the concept of evaluation mode and identifies it as a major determinant of evaluability: Provided that prior knowledge is low and attributes do not map onto an innate psychological scale, evaluability is higher when multiple options are judged than when one option is considered in isolation. This principle, taken jointly with GET's core postulate, predicts the EI—greater value sensitivity for poorly known attributes in joint than separate evaluation. Experiment 3 of the present article provides a counterexample to this prediction. "Value sensitivity" to an unknown index of sound distortion was lower within the {A, B, C, D} ("JE") condition than across the {A} and {B} ("SE") conditions. The options-as-information model accounts for this phenomenon by considering a range of individual contexts (in place of the joint-separate dichotomy) and a larger space of posterior models (in place of the unidimensional construct of evaluability).

Participants' behavior in these experiments was consistent with the normative model. However, it is important to note that the normative model only makes predictions about certain aspects of behavior (and not others). First, the model predicts that, for attributes that are poorly known, the parameters of the posterior model will be drawn toward the statistics of the option sample. The judgments of modelers in Experiment 1 clearly conform to this qualitative prediction. Importantly, however, the normative model does not specify the size of this assimilation effect. As noted earlier, the magnitude of the normative assimilation effect depends on the DM's assumptions about the process whereby options are sampled (e.g., whether sampling is representative or random) and on the details of the DM's task strategy (e.g., whether an average or a modal model is used; cf. footnote 1). Therefore, while the inferences of modelers in Experiment 1 are qualitatively reasonable, no strong claims can be made about their quantitative optimality. We cannot say whether participants employ a fully adequate algorithm in updating their beliefs, or, in particular, whether they would be appropriately sensitive to explicit information about the sampling process if it were provided.

The normative model does, however, make a strong additional prediction. It implies that, whatever inferences are drawn from different contexts, the effects of inferences on evaluations should match the effects of context. This, we have argued, is the appropriate test of coherence in the joint-separate design. Experiments 1 and 2 provide striking evidence that participants' evaluations are coherent in this sense.

The present work, like the study of JSRs more generally, is situated in a research tradition which identifies rationality with coherence. In this tradition, the theorist does not dictate substantive preferences to the DM, but requires only that judgments and choices be coherent. That is, evaluations elicited in varying ways should converge on a single underlying representation of value. Those who regard JSRs as irrational do so because the effects seem to violate this consistency requirement. However, this assessment overlooks an important qualification: Rational choice implies con-

sistency within, but not across, states of information. When different contexts lead to different inferences, the evaluations that result need not converge on a single representation of value. Notably, however, it *is* possible to examine coherence across information states—that is, to assess dynamic coherence when beliefs are fluid, rather than static coherence when beliefs are fixed. The assessment of dynamic coherence requires tailored research designs, which ask whether evolving beliefs and evolving preferences mutually cohere. The yoked modeler-recipient task reported here was designed to address this question. The results indicate that JSRs are dynamically coherent.

The remainder of this discussion is organized as follows. First, we situate our contribution in the broader literature on preference reversals and context effects. Next, we consider some conceptual and methodological limitations of the present analysis. We conclude with some comments on rational choice in context.

Scope

While the options-as-information model provides a natural account of one widely studied class of JSRs, it does not explain all apparent preference reversals. First, the model does not apply to effects of "response mode" rather than evaluation context. In these paradigms, the ordering of options reverses when different types of response are made—e.g., choice versus pricing (Lichtenstein & Slovic, 1971), other comparative versus noncomparative judgment scales (Nowlis & Simonson, 1997), and predicted experience versus decision (Hsee, 1999; Hsee, Zhang, Yu, & Xi, 2003). Because response mode effects are obtained even when the context of evaluation is constant, they do not reflect inferences from the set of options sampled.

Second, the model does not apply to JSRs that stem from incommensurability rather than ignorance—i.e., from the difficulties of trading off disparate social goods against one another and against private goods. For example, people may allocate more to environmental preservation when it is contemplated alone, but pay more to combat a human illness when the programs are directly compared (Kahneman & Ritov, 1994). Similarly, WTP for enhancement of a consumer good may exceed WTP for air quality improvement when these options are considered separately, even if people prefer air quality improvement when the options are considered jointly (Irwin et al., 1993). In these problems, the relevant options come from radically different categories, and the attributes of one option are therefore unlikely to impact the interpretation of the second option's attributes. Rather, these problems presumably arise because different options tap into different underlying metrics of psychological value (e.g., personal consumption vs. social welfare), and it is easier to scale options within metrics than to make trade-offs across metrics. Indeed, the problem of reconciling apparently incommensurable values (e.g., how many acres of restored habitat is a human life worth?) presents as formidable a challenge to ethicists (Chang, 1997), legal theorists (Adler, 1998), and economists (Sen, 1980) as it does to participants in decision making experiments. It falls beyond the scope of the present model.

We also note that simpler kinds of JSR may be generated with noninformational manipulations. In particular, as we explain in a footnote,[7] it should be possible to leverage *any* between-subjects effect of a transparently irrelevant variable to yield a JSR in a suitable design. In summary, while the options-as-information model provides a natural explanation for JSRs involving a single unfamiliar attribute, it does not purport to explain all preference reversals or all JSRs.

Going beyond JSRs, an options-as-information analysis may shed light on some regularity violations in the literature on context effects. The principle of regularity states that adding options to a choice set should not increase the popularity of any item in the original set. However, if attributes are poorly known, added options may convey new information about the underlying attribute distribution, and regularity violations can be rational. This suggestion is consistent with several models and findings in the literature. Ratneshwar, Shocker, and Steward (1987) suggested an informational interpretation of asymmetric dominance effects, in which an added "decoy" option draws preference share to an option that dominates it (Huber, Payne, & Puto, 1982). The decoy may suggest that the option that dominates it is relatively strong on both attributes (or that it features an unusually favorable cross-attribute trade-off). In a kindred development from economics, Kamenica (2008) has proposed an informational analysis of the emergence of context effects in market settings. Kamenica's analysis focuses on the compromise effect (Simonson, 1989) and choice overload (Iyengar & Lepper, 2000), and he remarks that similar considerations may apply to JSRs. In addition, as we noted earlier, an options-as-information analysis may shed light on several effects of global background context on choice (Mellers & Cooke, 1994; Stewart, 2009; Yeung & Soman, 2005), related to the preference reversal in Experiment 3. A fuller delineation of the scope and limits of sample-based inference in the broader literature on context effects is, we think, an important task for future research. Applying the yoked modeler-recipient design developed here to a range of experimental context effects might usefully contribute to that effort.

Limitations

This section considers limitations of the present model. We discuss idealizations in the model and then highlight some presuppositions of the methodology used to test it.

Model.   In modeling belief updating in the joint-separate design, we excluded inferences involving the easy-to-evaluate attribute. In principle, this attribute may figure in two types of inference. First, if this attribute it not *perfectly* known, DMs may draw weak inferences about its distribution. Second, if DMs believe that quality is positively correlated across attributes, values on the familiar attribute may reasonably influence interpretations of the corresponding values on the unfamiliar attribute.

These omissions conveniently simplified the model but are unlikely to affect its main predictions. First, to establish the EI,

---

[7] For example, suppose that, by associative priming, people are biased to assign higher benevolence ratings to the actions of a person named *Good*man rather than *Bad*deley. The actor's name is a transparently irrelevant variable that, by supposition, has a between-subjects effect on ratings. Now take Goodman's action, enhance its moral properties by an increment smaller than the priming effect, and attribute this marginally enhanced action to Baddeley. The natural prediction is that Goodman's deed will receive a higher average benevolence rating than Baddeley's when acts are evaluated in isolation, but Baddeley's action will receive a higher rating when the acts are considered jointly.

we require only that normative assimilation be far stronger for the difficult- than for the easy-to-evaluate attribute. This will be true even if there is nonzero assimilation for the latter. Second, inferences from cross-attribute correlations, while not necessary for the EI, may simply serve to reinforce it. For example, in the programmer problem, A's high GPA may suggest a better-than-average YT score when this candidate is seen in isolation, inflating the EI.

Whatever form they take, the inferences participants draw must operate on attended and remembered information only, subject to general capacity limits of attention and working memory (Cowan, 2001; Miller, 1956; Pashler, 1998). Because the standard joint-separate design involves unlimited viewing of a few attribute values, capacity limits are unlikely to seriously constrain inference in this case, and were not included in the present model. However, in designs that feature more attribute information and/or sequential rather than simultaneous presentation of options, capacity limits may become important. Thus the conditions of a joint-separate design can sometimes be implemented within-subjects, especially if the conditions are spaced out and interrupted with filler tasks to minimize memory carryover (González-Vallejo & Moran, 2001).[8]

Method. To test the model, Experiments 1 and 2 employed a modeler-recipient methodology. This design sought to approximate the critical thought experiment, in which the posterior worldview of one DM is transplanted into the mind of a second DM. The quality of the approximation depends on several important assumptions.

First, the descriptions provided by modelers—while necessarily crude and vague—must adequately capture the relevant aspects of their posterior view of the market. The efficacy of the design thus requires that these aspects have been identified and that the researcher has selected a question that targets them. Second, recipients must adequately attend to the (relatively pallid) background information. Third, recipients must take the background information at face value, even if (as in the A|{B} and B|{A} conditions of Experiment 2) it appears to conflict with the more vivid foreground evaluation set. Note that, in principle, recipients might discount the background information and incorporate the foreground information into their own posterior models.

The results of Experiments 1–2 clearly indicate that recipients attended to the background information they were given, and strongly suggest that they took it at face value. However, background information may not be taken at face value (or be adequately attended) if the foreground evaluation set is made larger, particularly when foreground and background information seem to conflict. Indeed, in an unpublished study in which recipients were confronted with a two-option rather than a single-option foreground evaluation set, we found only weak and erratic effects of background information.[9] These considerations suggest that the modeler-recipient design may form a good approximation to the critical thought experiment only when conflicting foreground information is minimized.

Rationality in Context

Rational choice theory assumes coherent preferences in a given information state. However, rational actors must be per-

mitted to learn—i.e., to move from one preference order to another as new information about the world is acquired. As a result, rational preferences and evaluations will generally be consistent within but not across states of information. When prior knowledge of attribute distributions is limited, the context of evaluation is potentially informative and context effects are potentially rational.

In joint-separate experiments, prior knowledge of one attribute is minimal by design. Participants are asked to form evaluations based on attributes, for one of which the sample of presented values is, in effect, the *only* information they have. In such circumstances, the best the participant can do is to construct an interpretation of the attribute on the basis of the contextual information, and form evaluations accordingly. However, informative contexts are not limited to the laboratory. In the wider world, the context of choice often serves both as a menu of options and a source of information.

Of course, contexts can be misleading. Even the best inference from the best available information will sometimes lead to inaccurate models and unfortunate choices. Furthermore, if interested parties can influence the choice context, even fully rational DMs are potentially exposed to manipulation (as long as the practice of manipulation is not so widespread as to render contexts totally uninformative). Vulnerability to error and deception is the price of ignorance. Fortunately, in the real world, it is usually possible, though sometimes costly, to obtain additional information. In that case, whether it is optimal to become an expert consumer or settle for smart comparison shopping depends on whether the price of information exceeds the price of ignorance. Ideally, we draw reasonable inferences in context when added information is not worth the cost, and we actively seek out fuller information when we cannot afford to go wrong.

However, participants in psychology experiments seldom have the latter option. The best they can do is to draw reasonable inferences from the options they have, and reasonably update their preferences in light of those inferences. The evidence reported here suggests that this is what people do in joint-separate experiments involving unfamiliar attributes. If so, judgments formed in different contexts can neither be conceptually combined nor normatively compared. That is, evaluation reversal does not imply preference reversal, and preference reversal does not imply irrationality.

---

[8] In general, attempts to approximate separate evaluation in within-subjects design have yielded mixed results (compare Experiments 1 and 3 in Willemsen & Keren, 2004). The feasibility of a within-subjects implementation may also be affected by whether the pragmatics of the experiment suggests that hypothetical scenarios in different conditions are to be construed as independent or as unfolding in a single connected world.

[9] In this study, recipients evaluated both options simultaneously based on background information derived from {A}, {B}, or {A, B} modelers. As in Experiment 2, we found that B was valued above A in all conditions. However, the source of the background information had no effect on valuations of A and B across the {A} and {B} conditions, in which the conflict between the foreground and background samples was greatest. For reasons that are unclear to us, WTP for B was lower relative to these conditions when background information was based on {A, B}, but the valuation of A did not change in this case.

References

Adler, M. (1998). Incommensurability and cost-benefit analysis. *University of Pennsylvania Law Review, 146,* 1371–1418. doi:10.2307/3312809

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology, 64,* 431–441. doi:10.1037/0022-3514.64.3.431

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14,* 471–485. doi:10.1017/S0140525X00070801

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior & Organization, 39,* 41–58. doi:10.1016/S0167-2681(99)00025-6

Beattie, J., & Baron, J. (1991). Investigating the effect of stimulus range on attribute weight. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 571–585. doi:10.1037/0096-1523.17.2.571

Birnbaum, M. H. (1999). How to show that $9 > 221$: Collect judgments in a between-subjects design. *Psychological Methods, 4,* 243–249. doi:10.1037/1082-989X.4.3.243

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Chang, R. (1997). *Incommensurability, incomparability, and practical reason.* Cambridge, MA: Harvard University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 114–185. doi:10.1017/S0140525X01003922

González-Vallejo, C., & Moran, E. (2001). The evaluability hypothesis revisited: Joint and separate evaluation preference reversal as a function of attribute importance. *Organizational Behavior and Human Decision Processes, 86,* 218–223. doi:10.1006/obhd.2001.2953

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14,* 357–364. doi:10.1016/j.tics.2010.05.004

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin, 118,* 248–271. doi:10.1037/0033-2909.118.2.248

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67,* 247–257. doi:10.1006/obhd.1996.0077

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making, 11,* 107–121. doi:10.1002/(SICI)1099-0771(199806)11:2<107::AID-BDM292>3.0.CO;2-Y

Hsee, C. K. (1999). Value seeking and prediction-decision inconsistency: Why don't people take what they predict they'll like the most? *Psychonomic Bulletin & Review, 6,* 555–561. doi:10.3758/BF03212963

Hsee, C. K. (2000). Attribute evaluability and its implications for joint-separate evaluation reversals and beyond. In D. Kahneman & A. Tversky (Eds.), *Choices, values, and frames* (pp. 543–563). Cambridge, England: Cambridge University Press.

Hsee, C. K., & Leclerc, F. (1998). Will products look more attractive when presented separately or together? *Journal of Consumer Research, 25,* 175–186. doi:10.1086/209534

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125,* 576–590. doi:10.1037/0033-2909.125.5.576

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5,* 343–355. doi:10.1177/1745691610374586

Hsee, C. K., Zhang, J., & Chen, J. (2004). Internal and substantive inconsistencies in decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 360–378). Oxford, England: Blackwell. doi:10.1002/9780470752937.ch18

Hsee, C. K., Zhang, J., Yu, F., & Xi, Y. (2003). Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making, 16,* 257–272. doi:10.1002/bdm.445

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90–98. doi:10.1086/208899

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. doi:10.1007/BF01065347

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79,* 995–1006. doi:10.1037/0022-3514.79.6.995

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus and Giroux.

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,* 5–37. doi:10.1007/BF01073401

Kamenica, E. (2008). Contextual inference in markets: On the information content of product lines. *American Economic Review, 98,* 2127–2149. doi:10.1257/aer.98.5.2127

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55. doi:10.1037/h0031207

List, J. A. (2002). Preference reversals of a different kind: The "more is less" phenomenon. *American Economic Review, 92,* 1636–1643. doi:10.1257/000282802762024692

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

MacLean, D. (2002). Some morals of a theory of nonrational choice. In R. Gowda & J. C. Fox (Eds.), *Judgments, decisions, and public policy* (pp. 46–70). Cambridge, England: Cambridge University Press.

McKenzie, C. R. M. (2003). Rational models as theories—not standards—of behavior. *Trends in Cognitive Sciences, 7,* 403–406. doi:10.1016/S1364-6613(03)00196-7

McKenzie, C. R. M. (2005). Judgment and decision making. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 322–339). London, England: Sage. doi:10.4135/9781848608177.n14

McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review, 10,* 596–602. doi:10.3758/BF03196520

Mellers, B. A., & Cooke, A. D. J. (1994). Trade-offs depend on attribute range. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1055–1067. doi:10.1037/0096-1523.20.5.1055

Mellers, B. A., & Cooke, A. D. J. (1996). The role of task and context in preference measurement. *Psychological Science, 7,* 76–82. doi:10.1111/j.1467-9280.1996.tb00333.x

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. doi:10.1037/h0043158

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research, 34,* 205–218. doi:10.2307/3151859

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. doi:10.1037/0033-295X.101.4.608

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418. doi:10.1037/h0022602

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications.* Mahwah, NJ: Erlbaum.

Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139173933

Ratneshwar, S., Shocker, A. D., & Steward, D. W. (1987). Toward understanding the attraction effect: The implication of product stimulus meaningfulness and familiarity. *Journal of Consumer Research, 13,* 520–533. doi:10.1086/209085

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105. doi:10.1037/0003-066X.54.2.93

Sen, A. (1980). Plural utility. *Proceedings of the Aristotelian Society, 81,* 193–215.

Shafir, E., & LeBoeuf, R. A. (2004). Context and conflict in multiattribute choice. In D. K. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 339–359). Oxford, England: Blackwell. doi:10.1002/9780470752937.ch17

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101,* 467–494. doi:10.1016/j.cognition.2005.11.001

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16,* 158–174. doi:10.1086/209205

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology, 62,* 1041–1062. doi:10.1080/17470210902747112

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26. doi:10.1016/j.cogpsych.2005.10.003

Wedell, D. H., & Pettibone, J. C. (1996). Using judgments to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes, 67,* 326–344. doi:10.1006/obhd.1996.0083

Willemsen, M. C., & Keren, G. (2004). The role of negative features in joint and separate evaluation. *Journal of Behavioral Decision Making, 17,* 313–329. doi:10.1002/bdm.476

Yeung, C. W. M., & Soman, D. (2005). Attribute evaluability and the range effect. *Journal of Consumer Research, 32,* 363–369. doi:10.1086/497547

Zikmund-Fisher, B. J., Fagerlin, A., & Ubel, P. A. (2004). "Is 28% good or bad?" Evaluability and preference reversals in health care decisions. *Medical Decision Making, 24,* 142–148. doi:10.1177/0272989X04263154