Organizational Behavior and Human Decision Processes 153 (2019) xxx–xxx

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

Beyond preference reversal: Distinguishing justifiability from evaluability in joint versus single evaluations

Xilin Li*, Christopher K. Hsee*

The University of Chicago Booth School of Business, United States

ARTICLE INFO

Keywords:
Decision-making
Preference reversal
Evaluability
Justifiability
Heuristics
Multi-attribute choice
Jury decision
Should-want conflict
Discrimination

ABSTRACT

Extensive existing research has studied how decisions differ between joint evaluation (JE) and single evaluation (SE), but most of the research aims to demonstrate preference reversals between two alternatives that vary on two attributes simultaneously. Thus, extant research cannot tell whether the reversal occurs because one of the attributes has a greater effect in JE than in SE, or the other attribute has a greater effect in SE than in JE, or both. Going beyond preference reversals, this research examines options that vary on only one attribute and studies whether the single attribute has a greater effect in JE or SE. We posit that any single attribute has two underlying characteristics—*evaluability* (i.e., whether people can evaluate a given value of the attribute without having to compare it with other values) and *justifiability* (i.e., whether people believe they should base their decisions on the attribute). Whether the single attribute has a greater effect in JE or SE depends on both the attribute's evaluability and justifiability. Specifically, (a) a high-justifiability/low-evaluability attribute (e.g., whether a candidate for a programming job has written 100 or 200 programs) has a greater effect in JE than in SE, and (b) a low-justifiability/high-evaluability attribute (e.g., whether the candidate belongs to a discriminated-against minority group) has a greater effect in SE than in JE. While the first proposition has been tested in prior research on evaluability, the second has not. Four experiments, including one in a naturally-occurring setting and another with orthogonal manipulation of evaluability and justifiability, tested and supported these propositions, especially the second.

1. Introduction

This research studies the relationship between two of the most basic aspects of decisions—attribute and evaluation mode. An *attribute* is a factor which differentiates between alternatives and about which the decision maker cares (i.e., finds relevant or is otherwise tempted to consider). *Evaluation mode* is how a decision is made; every decision is made in either the joint evaluation (JE) mode (in which different alternatives are presented and evaluated simultaneously), the single evaluation (SE) mode (in which each alternative is presented and evaluated in isolation), or some combination of the two. (Readers should not confuse difference in evaluation mode with difference in response type, such as choice versus rating. In all of our experiments, we held response type constant and varied only evaluation mode. For more information on how evaluation mode differs from response type, see Hsee, Loewenstein, Blount, & Bazerman, 1999.)

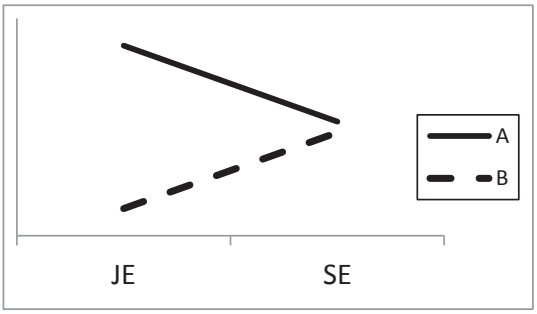
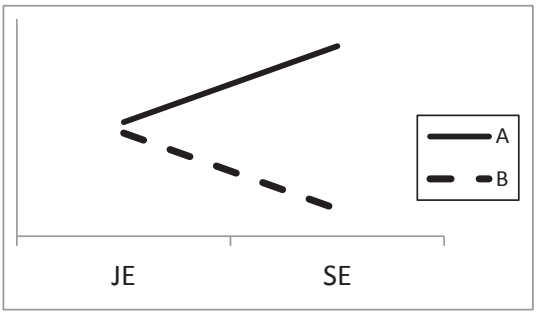
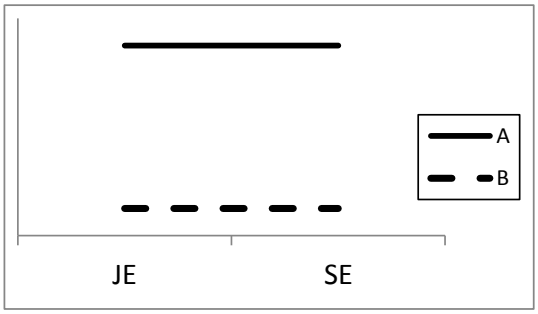
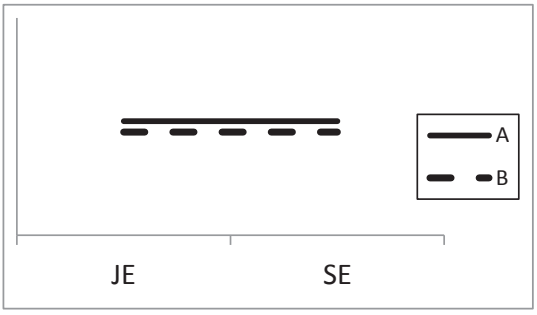
More specifically, this research studies whether a given attribute (e.g., the race of a job candidate) has a greater effect in JE or SE (e.g., makes a greater difference in job recruiters' assessments of the candidates when the candidates are evaluated jointly or separately), while holding other attributes constant. Understanding this question can not only enrich our knowledge of basic human judgment and decision processes, but also help the relevant party make better choices. For example, suppose that you belong to a discriminated-against ethnic minority group. You are applying for a job, and you have the option to be interviewed alone or jointly with another candidate who is similar to you except that he is not a minority. You can make a savvier decision if you understand whether ethnicity has a greater effect in JE or SE.

Before continuing, let us first introduce several terms that we will use throughout this article. We use "the effect of an attribute" to mean the *difference* in evaluation between two alternatives that differ only on this attribute (e.g., the difference in assessment between two job candidates who differ only in race). We use "JE amplification" to mean a greater difference in JE than in SE and "SE amplification" to mean a greater difference in SE than in JE. See the top two panels of Table 1 for illustrations of JE amplification and SE amplification. Note that both JE amplification and SE amplification are about interactions between evaluation mode and alternatives. We are not concerned with main

* Corresponding authors at: Booth School of Business, University of Chicago, 5807 South Woodlawn Avenue, Chicago, IL 60637, United States (C.K. Hsee).
E-mail addresses: xilin.li@chicagobooth.edu (X. Li), Chris.Hsee@ChicagoBooth.edu (C.K. Hsee).

https://doi.org/10.1016/j.obhdp.2019.04.007
Received 29 March 2018; Received in revised form 21 April 2019; Accepted 25 April 2019

Table 1
Summary of propositions. (Each graph shows how the responses to two alternatives, A and B, differ between JE and SE.)

Proposition	Justifiability	Evaluability	Relative effect between JE and SE	Stylized illustration
1	High	Low	More effective in JE, i.e., *JE amplification*	
2 (main proposition)	Low	High	More effective in SE, i.e., *SE amplification*	
3	High	High	Effective in both JE and SE	
4	Low	Low	Ineffective in both JE and SE	

effects of evaluation mode. For discussions on JE-SE main effects, see Hsee and Leclerc (1998).

2. Existing research

A large body of literature has studied how preferences are constructed and how they vary across different elicitation procedures (e.g., Lichtenstein & Slovic, 2006; Mellers, Chang, Birnbaum, & Ordonez, 1992; Nowlis & Simonson, 1997). Within that body of literature, extensive research has studied how decisions differ between JE and SE. However, most existing research on JE versus SE aims to demonstrate JE-SE preference reversals between alternatives that involve a tradeoff between two attributes (e.g., Bazerman, Tenbrunsel, & Wade-Benzoni, 1998; González-Vallejo & Moran, 2001; Hsee, 1996a; Hsee, 1998; Hsee, Zhang, Wang, & Zhang, 2013; Shaffer & Arkes, 2009; Montanari, Thomas, Treyball, & Acikgoz, 2017). For example, Hsee (1996a) asked research participants at the University of Illinois in Chicago (UIC), which had a 5-point GPA system, to evaluate one or both of two job candidates, who presented a tradeoff between two attributes: programming experience and GPA.

	Candidate A	Candidate B
Programming experience	Has written 70 programs	Has written 10 programs
GPA from UIC	3.0	4.9

The study revealed a JE-SE preference reversal: in JE, Candidate A was evaluated more favorably, but in SE, Candidate B was evaluated more favorably. Such JE-SE preference reversals have been replicated in many domains.

Because this type of research aims to demonstrate JE-SE preference reversals and uses alternatives that vary on two attributes simultaneously, it cannot tell whether the reversal is due to one attribute having a greater effect in JE than in SE (JE amplification), or the other attribute having a greater effect in SE than in JE (SE amplification), or both. Nor can it address a more basic question: holding other attributes constant, will a single attribute show JE amplification or SE amplification? The present research is trying to address this more basic question rather than to demonstrate JE-SE preference reversals.

Although most existing studies have used alternatives that vary on two attributes simultaneously, a few existing studies have focused on alternatives that vary on only one attribute. These studies have typically found JE amplification (e.g., Dunn, Koehler, & Risko, 2017; Hsee & Zhang, 2004; Hsee et al., 2013; Krüger, Mata, & Ihmels, 2014). For example, a study assessing people's support for a bear-rescue program found that the number of bears saved—100 versus 200—showed JE amplification. In other words, this attribute made a greater difference in people's support in JE than in SE (Hsee et al., 2013).

Hsee and colleagues (e.g., Hsee, 1996a; Hsee et al., 1999; Hsee & Zhang, 2010) ascribe JE amplification to low evaluability. Evaluability refers to the extent to which people can evaluate a given value of the attribute without having to compare it with other values. According to the notion of evaluability, most attributes show some degree of JE amplification; low-evaluability attributes show more JE amplification than high-evaluability attributes because low-evaluability attributes require JE to differentiate between the attributes' values, whereas high-evaluability attributes do not. For example, according to the evaluability hypothesis, in the aforementioned programmer study, the number of programs a job candidate had written (in this case, 70 versus 10) was a low-evaluability attribute and showed JE amplification because, without a comparison, most participants could not discern the quality of the given number of programs and only in comparison did the numbers make sense. By contrast, the GPA of a job candidate (in this case, 3.0 versus 4.9) was a high-evaluability attribute and did not show much JE amplification; namely, it was similarly influential in JE and in SE, because even without a comparison, the participants understood the quality of a given GPA.

Evaluability can explain JE-SE preference reversals for alternatives that involve a tradeoff between two attributes. Consider two alternatives, one superior on a low-evaluability attribute and one superior on a high-evaluability attribute. Then, as long as the difference between the alternatives on the low-evaluability attribute is large enough, the preference ordering of the two alternatives will reverse between JE and SE. Specifically, the alternative that is superior on the low-evaluability attribute will be favored in JE, and the alternative that is superior on the high-evaluability attribute will be favored in SE. This analysis explains why the programmer study (Hsee, 1996a) found that the job candidate superior on programming experience (a low-evaluability attribute) was favored in JE, while the job candidate superior on GPA (a high-evaluability attribute) was favored in SE.

Notably, a JE-SE preference reversal requires only one attribute (e.g., work experience) to show more JE amplification than the other (e.g., GPA), and does not require either attribute to show SE amplification. In other words, the presence of a JE-SE preference reversal does not imply that either attribute shows SE amplification. See Appendix A for an illustration.

To summarize, the existing literature on the evaluability of single attributes has found JE amplification (with more JE amplification for low-evaluability attributes than high-evaluability attributes), but never SE amplification.

3. Current research

Can an attribute ever show SE amplification—having a greater effect in SE than in JE while holding other attributes constant? To address the question, we introduce the notion of justifiability. Justifiability is independent of evaluability; it refers to the extent to which people believe they should consider the attribute in their decision in the given context (e.g., Lerner & Tetlock, 1999).

Justifiability and evaluability are two of the most general characteristics underlying all attributes. Any specific attribute can be described as high or low on evaluability, and high or low on justifiability. Whether an attribute is high or low on justifiability is context-dependent. In most contexts, decision makers know whether a given attribute is of high justifiability or low justifiability. For example, in most job recruitment contexts, recruiters know that the work experience and GPA of a job candidate are high-justifiability attributes—factors on which they should base their decision—while the race and gender of a job candidate are low-justifiability attributes—factors on which they should not base their decision. In this research, we do not define the justifiability of an attribute arbitrarily; rather, we operationally define an attribute to be of high or low justifiability based on whether pretest participants think they should or should not consider the attribute in the given decision context.

Previous research on evaluability has not distinguished evaluability from justifiability and has focused only on high-justifiability factors, such as work experience and GPA. The present research focuses on low-justifiability attributes and proposes that, holding evaluability at a high level, low-justifiability attributes show SE amplification. This proposition builds on extant research showing that low-justifiability attributes are more likely to influence decision makers if there is ambiguity than if there is not (Darley & Gross, 1983; Hsee, 1995, 1996b; Norton, Vandello, & Darley, 2004; Snyder, Kleck, Strenta, & Mentzer, 1979). For example, the race of a college applicant is more likely to influence admission decisions if there is ambiguity in the applicant's academic records than if there is no ambiguity (Hodson, Dovidio, & Gaertner, 2002). Although this stream of research neither concerns nor manipulates evaluation mode, it suggests that with everything else equal, low-justifiability attributes are more influential in SE than in JE, because SE generally entails more ambiguity than JE.

For example, suppose that two candidates are being evaluated for a job, and the job recruiters are supposed to consider the candidates' work experience, only—and no other attributes. Suppose also that the two candidates have the same work experience, but one is white and the other is (non-white) Hispanic. The two job candidates may be evaluated in either JE or SE. We predict that the two job candidates will receive similar evaluations in JE, because JE makes it clear that the two candidates have the same work experience; the recruiters will have little justification—either to themselves or to other people—to give the white candidate a more favorable evaluation. On the other hand, we predict that the white candidate will be evaluated more favorably than the Hispanic candidate in SE, because SE masks the fact that the two candidates have identical work experience and thereby affords leeway for the race of the job candidates to inform the recruiters' evaluations, either consciously or unconsciously (Hodson et al., 2002; Hsee, 1996b; Norton et al., 2004; Snyder et al., 1979). In short, we argue that JE inhibits the influence of a low-justifiability attribute, while SE facilitates it. Consequently, a low-justifiability attribute shows SE amplification, namely, a greater effect in SE than in JE.

We are not the first to make the above proposition. Bazerman and colleagues have drawn a distinction between "should" attributes (attributes that decision makers think they *should* consider) and "want" attributes (attributes that decision makers *want* to consider), and posited that "should" attributes are more influential in JE, while "want" attributes are more influential in SE (Bazerman et al., 1998; Bitterly, Mislavsky, Dai, & Milkman, 2014; Bohnet, Van Geen, & Bazerman, 2016; Hsee et al., 1999). This latter proposition has inspired our hypothesis that low-justifiability/high-evaluability attributes show SE amplification.

However, the present research extends the existing research on "should" versus "want" in several ways. First, prior work on "should" and "want" is concerned with justifiability and not with evaluability. By

contrast, the present research is concerned with both. Second, "should" and "want" are not two ends of a single dimension, because a want-attribute can also be a should-attribute, for example, for most car buyers, the price of a car is both a want-attribute and a should-attribute. By contrast, low-justifiability and how-justifiability are two ends of a single dimension. Finally and most importantly, even if we assume that "want" attributes are equivalent to low-justifiability attributes, there is no solid evidence for Bazerman and colleagues' proposition that "want" attributes are more influential in SE than in JE. The best available evidence for the proposition comes from a study conducted by Bohnet et al. (2016). The authors attributed the result of the study to differential justifiability, but the result could alternatively be explained by differential evaluability. The study examined students' preferences for two candidates to work on a stereotypically gender-specific task, such as a math task, for which men are stereotypically considered better. The two candidates presented a tradeoff between two attributes: (a) past performance (i.e., the candidate's number of correct answers on the task in the past) and (b) gender. For example, in the math-task condition, the two candidates were as follows:

	Candidate A	Candidate B
Past performance	Had 10 correct answers	Had 9 correct answers
Gender	Female	Male

The research participants displayed a JE-SE preference reversal: in JE, the participants preferred Candidate A, and in SE, the participants preferred Candidate B. A similar JE-SE preference reversal occurred for a stereotypically female-superior task. Presumably, gender was a "want" (low-justifiability) attribute, and the observed JE-SE preference reversal could be interpreted as evidence that a low-justifiability attribute showed SE amplification.

However, the JE-SE preference reversal demonstrated by Bohnet et al. could also be explained in terms of evaluability. The structure of this study was very similar to that of the aforementioned programmer study (Hsee, 1996a): Both programming experience and past performance on a math task are of low evaluability, while both GPA and gender are of high evaluability. Thus, like the result of the programmer study, the result of this study may have occurred because the low-evaluability attribute (past performance) showed more JE amplification than did the high-evaluability attribute (gender). (Although participants in the Bohnet et al. study were told about the average past performance, the task was unfamiliar, and relatively speaking, past performance on the task was still of lower evaluability than gender.)

To be clear, we are not arguing that gender is a high-justifiability attribute. Rather, our point is that even if gender is a low-justifiability attribute, the preference reversal reported by Bohnet and colleagues can still be explained parsimoniously by evaluability, and therefore cannot be treated as conclusive evidence for the proposition that a low-justifiability attribute shows SE amplification while holding other attributes constant. Indeed, even the authors of that study themselves were unsure whether gender would show SE amplification while holding other attributes constant; they noted, "(if) gender is the only variable that differs in this condition, with everything else being identical, gender is likely more salient than in separate evaluation and even than in our existing joint-evaluation condition with mixed-sex and mixed-performance-level pairs. Gender salience may either lead to an increase in stereo-typical choices or to reactance and a decrease in stereotypical choices, thus truly making this an empirical question beyond the scope of this paper" (Bohnet et al., 2016, p. 1229).

In short, even though various existing lines of research suggest that low-justifiability attributes may show SE amplification, most of the evidence comes from preference reversals involving multiple attributes. As noted earlier, in order to test whether a given attribute shows SE amplification, researchers should not use alternatives that vary on multiple attributes and should not look for JE-SE preference reversals, which inherently require alternatives that vary on multiple attributes. Rather, researchers should use alternatives that vary on only one attribute and test whether that attribute commands a greater effect in SE than in JE. This is the structure of the present research. In all of our studies, we manipulate only one attribute while holding everything else constant, and we test whether that attribute shows SE amplification.

4. Summary of propositions and overview of experiments

4.1. Propositions

This research studies whether a single attribute shows JE amplification or SE amplification while holding other attributes constant. We posit that the answer depends on two underlying characteristics of the attribute—evaluability and justifiability. Specifically, we propose the following:

First, if an attribute is high in justifiability but low in evaluability, it will be more influential in JE than in SE, namely, it will produce a *JE amplification effect*. Second, if an attribute is low in justifiability but high in evaluability, it will be more influential in SE than in JE, namely, it will produce an *SE amplification effect*. Third, if an attribute is high in both justifiability and evaluability, it will be influential in both JE and SE. Finally, if an attribute is low in both justifiability and evaluability, it will be uninfluential in both JE and SE. See Table 1 for a summary of the propositions.

Note that of the above propositions, only the first two make a directional prediction regarding the relative effect of an attribute between JE and SE. And, as discussed earlier, of these two propositions, the first proposition (JE amplification of high-justifiability/low-evaluability attributes) has already been tested in prior research on evaluability, but the second proposition (SE amplification of low-justifiability/high-evaluability attributes) has never been tested cleanly before. Thus, the primary purpose of this research is to test this SE-amplification proposition.

4.2. Overview of experiments

The ensuing sections report four experiments that covered different contexts, used both US and Chinese participants, and involved both scenario-based and real-life decisions. Experiment 1A and Experiment 1B were about jury decisions; Experiment 1A involved a high-justifiability/low-evaluability attribute (the number of victims) and sought to replicate the JE amplification effect demonstrated in previous research, while Experiment 1B involved a low-justifiability/high-evaluability attribute (the ethnicity of the victims) and sought to test the SE-amplification proposition (our main proposition).

Experiment 2 was a field experiment concerning a tutoring service; like Experiment 1B, Experiment 2 focused on a low-justifiability/high-evaluability attribute (the ethnicity of the tutee) and sought to demonstrate that the SE amplification effect occurs not only in hypothetical scenarios, but also in naturally-occurring settings.

Finally, Experiment 3 was about job candidate evaluations; it directly and orthogonally manipulated the justifiability and evaluability of a single attribute (the test score of a job candidate) and tested all propositions of our theory. See Table 2 for an overview of the experiments.

5. Experiments 1A and 1B

The two experiments were similar except that Experiment 1A involved a low-evaluability/high-justifiability attribute and tested the JE-amplification proposition, and Experiment 1B involved a high-evaluability/low-justifiability attribute and tested the SE-amplification proposition.

Table 2
Overview and main results of experiments.

Experiment	Attribute	Justifiability	Evaluability	Evaluation mode	Alternative	DV	Effect, i.e., difference in DV between the two alternatives	Relative effect between JE and SE, i.e., interaction between evaluation modes and alternatives
1A (DV = jail term in years)	Victim number	High	Low	JE	20 victims	6.58	1.20***	JE amplification*
					10 victims	5.38		
				SE	20 victims	4.21	−0.12	
					10 victims	4.33		
1B (DV = jail term in years)	Victim ethnicity	Low	High	JE	Belgian	3.95	0.04	SE amplification***
					Somalian	3.91		
				SE	Belgian	4.60	1.92***	
					Somalian	2.68		
2 (DV = tutor fee in $)	Tutee ethnicity	Low	High	JE	French	11.82	0.14	SE amplification*
					Guinean	11.68		
				SE	French	9.27	−4.44*	
					Guinean	13.71		
3 (DV = evaluation of job candidate on a 100-point scale)	MED Test score	High	Low	JE	High score	91.35	34.35***	JE amplification***
					Low score	57.00		
				SE	High score	79.21	20.95***	
					Low score	58.26		
		Low	High	JE	High score	78.83	5.37*	SE amplification*
					Low score	73.46		
				SE	High score	75.41	14.18***	
					Low score	61.23		
		High	High	JE	High score	96.33	77.52***	No significant difference
					Low score	18.81		
				SE	High score	90.22	74.44***	
					Low score	15.78		
		Low	Low	JE	High score	79.84	1.82	No significant difference
					Low score	78.02		
				SE	High score	72.70	0.28	
					Low score	72.42		

* $p < .05$.
*** $p < .005$; for more detailed statistics, see main text.

5.1. Method

Both Experiments 1A and 1B included three between-participants evaluation-mode conditions, one in JE and two in SE. We recruited participants from MTurk in the US and set a target sample size of 255 (85 per cell) per experiment. We received completed responses from 258 participants (157 females; $M_{age} = 34.42$, $SD_{age} = 11.40$) for Experiment 1A and 240 participants (142 females; $M_{age} = 32.64$, $SD_{age} = 11.30$) for Experiment 1B.

In each experiment, we first told the participants that we wanted to know their own opinions and that we would keep their responses anonymous and confidential. We then asked participants to assume the role of jurors in a case involving a US fighter pilot who had killed civilians overseas. Participants in the JE condition read two alternative scenarios, and participants in each of the two SE conditions read just one of the two scenarios. In Experiment 1A, the attribute that differentiated the two scenarios was the number of victims (10 or 20); in Experiment 1B, the attribute that differentiated the two scenarios was the ethnicity of the victims (Belgian or Somalian). The number of victims was a low-evaluability attribute because the killing of civilians overseas is an unusual event; without a direct comparison, participants had little basis to gauge the severity of 10 civilians versus 20. By contrast, the ethnicity of the victims was a high-evaluability attribute; extant research suggests that even without direct comparison, whites or Europeans evoke more positive attributes than do blacks or Africans (e.g., Nosek, Banaji, & Greenwald, 2002). Furthermore, we assumed, and verified in a pretest (see below), that the number of victims was a high-justifiability attribute and the ethnicity of victims was a low-justifiability attribute.

Specifically, in Experiment 1A, participants assigned to the JE condition read a questionnaire with two alternative scenarios, which were identical except for the number of victims involved:

> *Consider two alternative scenarios:*
> *Scenario A: In a recent overseas military operation, a US fighter pilot mistakenly fired a missile and killed 10 Japanese civilians in Asia.*
> *Scenario B: In a recent overseas military operation, a US fighter pilot mistakenly fired a missile and killed 20 Japanese civilians in Asia.*
> *The fighter pilot is now on trial for his behavior, and you are a juror. Legally, the punishment can range anywhere from 0 years to 10 years of imprisonment. How many year(s) of imprisonment would you recommend in each scenario?*

The participants selected a number between 0 and 10 for each scenario. The SE conditions were identical to the JE condition except that each participant read only one of the scenarios and decided on the sentencing term for that scenario.

Experiment 1B was structured the same way as Experiment 1A, except that the scenarios varied in the ethnicity of the victims:

> *Scenario A: In a recent overseas military operation, a US fighter pilot mistakenly fired a missile and killed 18 Somalian civilians in Africa.*
> *Scenario B: In a recent overseas military operation, a US fighter pilot mistakenly fired a missile and killed 18 Belgian civilians in Europe.*

5.2. Results and discussion

5.2.1. Justifiability pretest

To verify the assumption that the number of victims was a high-justifiability attribute and the ethnicity of victims was a low-justifiability attribute, we described the fighter-pilot case to a group of pretest participants drawn from the same pool as in the main experiment ($N = 50$; 29 females; $M_{age} = 34.06$, $SD_{age} = 11.73$). We asked participants to rate, on a 5-point scale (1 = “should not consider”; 5 = “should consider”), the extent to which they believed they should

consider each of the two factors. Confirming our assumptions, the pretest participants rated the number of victims above the midpoint of the scale ($M = 3.58$, $SD = 1.49$, $t(49) = 2.76$, $p = .008$) and the ethnicity of the victims below the midpoint ($M = 1.24$, $SD = 0.72$, $t(49) = 17.38$, $p < .001$).

5.2.2. Main results of Experiment 1A

We predicted that the attribute in Experiment 1A—the number of victims—would show JE amplification. To test the prediction, we needed to determine whether there was a significant interaction between evaluation modes (JE versus SE) and alternatives (10 victims versus 20 victims). However, because the two alternatives were presented within participants in JE and between participants in SE, we could not run a regular ANOVA to test the interaction. Instead, we used a contrast code regression analysis to test the interaction, and we describe the code in detail in Appendix B. Supporting our prediction and replicating the previous findings in the evaluability literature, the analysis revealed a significant interaction effect ($\beta = -0.26$, $p = .02$; see Table 2). Specifically, in JE, participants imposed a significantly more severe punishment if the pilot had killed 20 civilians ($M = 6.58$, $SD = 3.60$) than if he had killed 10 civilians ($M = 5.38$, $SD = 3.63$), $t(83) = 6.66$, $p < .001$. In SE, participants awarded similar sentencing terms regardless of whether the pilot had killed 20 civilians ($M = 4.21$, $SD = 3.57$) or 10 civilians ($M = 4.33$, $SD = 3.53$), $t(172) = 0.22$, $p = .83$. (We note in passing that on average, punishments made in JE were harsher than punishments made in SE. This is a JE-SE main effect, and, as noted earlier, is beyond the scope of this research.)

5.2.3. Main results of Experiment 1B

We predicted that the attribute in Experiment 1B—the ethnicity of the victims—would show SE amplification. Confirming the prediction, the result revealed a significant interaction between evaluation modes and alternatives ($\beta = 0.43$, $p < .001$, using the same test described above). Specifically, in JE, participants imposed similar jail terms on the pilot regardless of whether he had killed Belgians ($M = 3.95$, $SD = 3.35$) or Somalians ($M = 3.91$, $SD = 3.38$), $t(81) = 0.40$, $p = .688$. But in SE, participants awarded a much lighter punishment—by almost 2 years—if the victims were Somalian ($M = 2.68$, $SD = 2.71$) than if the victims were Belgian ($M = 4.60$, $SD = 3.46$), $t(156) = 3.85$, $p < .001$.

5.2.4. Discussion

Both Experiments 1A and 1B used alternatives that varied on only one attribute. In Experiment 1A, the attribute was high in justifiability and low in evaluability, and it showed JE amplification. This result was a replication of the JE amplification effect already shown in previous studies (e.g., Hsee et al., 2013). By contrast, in Experiment 1B, the attribute was low in justifiability and high in evaluability, and it showed SE amplification. To the best of our knowledge, this is the first study to demonstrate such an effect cleanly.

Experiments 1A and 1B also carry practical implications for jury decisions, suggesting that the same two cases could be judged rather differently depending on whether they are presented and evaluated jointly or separately. Furthermore, the direction of the difference depends on whether the cases differ on a low-evaluability attribute or on a low-justifiability attribute. The finding of Experiment 1B, regarding the ethnicity of the victims, is particularly noteworthy. Scholars and practitioners have debated about the prevalence of ethnic discrimination in legal decisions (Abrams, Bertrand, & Mullainathan, 2012). Our study suggests that jurors exhibit little discrimination in JE, but significant discrimination in SE.

6. Experiment 2

Experiment 1B provided initial evidence for the SE amplification proposition regarding low-justifiability/high-evaluability attributes, showing that the ethnicity of airstrike victims exerted greater influence on a jury verdict made in SE than in JE. To test the external validity and the generality of the finding, Experiment 2 sought to replicate the finding in a real-life decision rather than a scenario-based decision, in a work-related domain rather than a legal domain, and with Chinese participants rather than US participants. The experiment examined how the ethnicity (French or Guinean) of a foreign student seeking a Chinese tutor would influence the prospective tutor's proposed rate; we predicted that the tutee's ethnicity would make a greater difference in SE than in JE.

6.1. Method

The experiment consisted of one JE and two SE conditions. Over a predetermined 10-day period, we placed an advertisement on a Chinese service-exchange website (www.zbj.com). The advertisement said that some foreign students were looking for Chinese language tutors, and it invited interested individuals to click a link to see more information. Unbeknownst to the visitors, the link led to one of three different postings, corresponding to the three experimental conditions.

According to the JE posting, two foreign students—one from France and one from Guinea—would visit China in the next few months, and they both wanted to hire a Chinese tutor to practice oral Chinese via the internet for four weeks before their trip. The posting asked interested individuals to indicate their price (a minimum hourly rate) for tutoring each student, and to leave their contact information. Each of the two SE postings was identical, except that they mentioned either the French student or the Guinean student, but not both.

During the 10-day period of the experiment, a total of 141 visitors responded to our advertisement (81 females; $M_{age} = 25.61$; $SD_{age} = 4.70$). After the experiment, we sent the participants a debriefing message.

6.2. Results and discussion

6.2.1. Justifiability pretest

We assumed that most people would consider the ethnicity of the tutee a low-justifiability attribute. To verify the assumption, we described the task in the main experiment to a group of pretest participants recruited from the same pool as in the main experiment ($N = 54$; 31 females; $M_{age} = 32.76$, $SD_{age} = 8.13$), promised them anonymity and confidentiality, and asked whether they should consider the ethnicity of the tutee, among other factors, when deciding how much to charge her. Supporting our assumption, the majority (77.8%) said no ($\chi 2$ (1, $N = 54$) = 16.67, $p < .001$, compared with 50%).

6.2.2. Main results

Even though ethnicity was pretested to be a low-justifiability attribute, it nevertheless produced a significant effect in SE, but not in JE. As Table 2 summarizes, a significant evaluation-mode × alternative interaction effect emerged from the responses to our advertisement, $\beta = -0.36$, $p = .026$, using the aforementioned contrast code. In JE, the Chinese participants showed little discrimination; they set similar prices (hourly rates, which we converted from Chinese renminbi [CN¥] to US dollars [$] at the exchange rate of CN¥6.50 = $1.00) for the French tutee ($M = 11.82$, $SD = 9.90$) and the Guinean tutee ($M = 11.68$, $SD = 9.96$), $t(45) = 1.26$, $p = .216$. But in SE, the participants exhibited significant discrimination; they set considerably lower prices for the French tutee ($M = 9.27$, $SD = 7.95$) than for the Guinean tutee ($M = 13.71$, $SD = 11.75$), $t(93) = -2.20$, $p = .031$, suggesting that they were more willing to tutor the former than the latter.

6.2.3. Discussion

The above analysis is conservative because it is based only on the people who responded to our advertisement. Due to technical constraints, we could not track how many people saw our advertisement

but left without responding. We know, however, that fewer people responded in the Guinean SE condition (41 people) than in the French SE condition (54 people), which suggests that more people in the Guinean SE condition may have left without responding. Had we been able to incorporate this difference in our analysis, the discrimination effect found in SE may have been even stronger than reported above.

This experiment is a conceptual replication of Experiment 1B. Instead of studying how US jury-qualifying participants make jury decisions involving victims of different ethnicities, this experiment studied how prospective Chinese service providers make pricing decisions for service-seekers of different ethnicities. Instead of relying on hypothetical scenarios, this experiment took place in a naturally-occurring setting, in which the participants did not know they were participating in an experiment. Despite the differences, this experiment yielded the same pattern of results as in Experiment 1B.

7. Experiment 3

Experiment 3 was primarily for theory testing, and it differed from the other experiments in several aspects. First, while Experiment 1A involved only a high-justifiability/low-evaluability attribute, and both Experiment 1B and Experiment 2 involved only a low-justifiability/high-evaluability attribute, Experiment 3 involved attributes that covered all four combinations of justifiability and evaluability. Second, while the other experiments compared the participants' responses to attributes that were qualitatively different (e.g., number of victims as the high-justifiability/low-evaluability attribute and ethnicity as the low-justifiability/high-evaluability attribute), Experiment 3 used the same attribute (the test score of a job candidate) for all four combinations of justifiability and evaluability.

Third, in each of the other experiments, we assumed (rather than manipulated) whether the attribute was high or low in justifiability and high or low in evaluability, but in Experiment 3, we directly and orthogonally manipulated the attribute's justifiability and evaluability. Fourth, in each of the other experiments, we assessed the justifiability of the attribute by asking pretest participants whether they should or should not consider the attribute in their decision, but in Experiment 3, we explicitly told participants whether they should or should not consider the attribute in their decision. This forthright manipulation reduced the potential ambiguity of justifiability.

Finally, while the other experiments implemented SE between participants (i.e., different participants evaluated the two alternatives), Experiment 3 implemented SE within participants (i.e., the same participants evaluated the alternatives at different times, separated by fillers). This design made the evaluation-mode manipulation both subtler and cleaner.

7.1. Method

The experiment comprised four between-participants attribute-type conditions: high-justifiability/low-evaluability, low-justifiability/high-evaluability, high-justifiability/high-evaluability, and low-justifiability/low-evaluability. Each of the attribute-type conditions included two between-participants evaluation-mode conditions: JE and SE. In other words, the study followed a 2 (justifiability, between-participants) × 2 (evaluability, between-participants) × 2 (evaluation mode, between-participants) × 2 (alternative, within-participants) mixed design.

We recruited participants from MTurk in the US. Following Experiments 1A and 1B, and recognizing that the manipulations in Experiment 3 were more direct, we set a target sample size of 520 (65 per cell). We received completed responses from 521 participants (285 females; M_{age} = 34.78, SD_{age} = 11.41). In all the conditions, participants were asked to rate several candidates from a foreign country who were applying for a paramedic job, presumably in the US. For each candidate, we provided information on multiple attributes, including gender, highest education, earliest starting date, and the candidate's score on a MED Test, which was described as a test in the candidate's country of origin.

Among the job candidates, only two were our stimuli, and the others were fillers. The two stimulus candidates differed on only one attribute—their MED Test score. We manipulated the evaluability of the score by using either numbers or words. In the low-evaluability condition, we described the score of one candidate as 900, and that of the other candidate as 1400. In the high-evaluability condition, we described the score of one candidate as "excellent," and that of the other candidate as "poor."

We manipulated the justifiability of the score as follows. In all conditions, we told participants in advance that they should base their evaluation of the candidates on only some of information (attributes) we gave them, and to ignore the rest. In the high-justifiability condition, the key attribute—the MED Test score—was among the information we told participants that they should consider, and in the low-justifiability condition, the key attribute was among the information we told participants they should ignore. We intentionally described the MED Test as a foreign test, because both telling participants to consider a foreign score and telling them to ignore a foreign test score sounded realistic.

We manipulated the evaluation mode as follows. In the JE condition, we showed participants the two stimulus candidates side by side on the same page, and we asked participants to evaluate each candidate by moving a slider on a scale anchored by 0 (not very qualified) to 100 (extremely qualified). In the SE condition, we showed participants the two job candidates on different pages, separated by a filler candidate and several filler studies. The filler candidate was identical to the first stimulus candidate except for gender (male instead of female). The filler studies were short questionnaires on unrelated topics, such as personal preference in friend-making and risk-preference in investment. We counterbalanced the order of the two stimulus candidates and found no meaningful systemic order effects. In all conditions, after reading the instructions but before evaluating the candidates, participants answered three manipulation-check/comprehension questions (see Appendix C).

As an example, participants assigned to the JE condition within the low-justifiability/high-evaluability condition read the following:

> *In this study, we will ask you to rate the qualifications of two candidates from a foreign country for a paramedic job. For each candidate, we will give you multiple pieces of information, such as gender, earliest starting date, and score on MED Test (an exam for paramedics in the foreign country). Your goal is to give an overall rating for the candidate.*
> *Please note:*
> *(1) We ask different participants to use different pieces of information when giving their overall ratings. For you, we want you to give your rating by using only the following pieces of information, and ignoring the other pieces of information. In other words, your rating of the candidate should be only based on the following two pieces of information:*
> *– Highest degree*
> *– Earliest starting date*
> *(2) You may assign the same overall rating or different overall ratings to the two candidates.*
> *[Manipulation-check/comprehension questions; see Appendix C]*
> *[Next page]*
> *Now, please start to rate the candidates.*

	Candidate H	*Candidate L*
Gender	*Female*	*Female*
Highest education	*Bachelor*	*Bachelor*
Earliest starting date	*May 1st*	*May 1st*
MED Test score	*Excellent*	*Poor*

How qualified is each candidate?
[Next page]
[Filler candidates, same as the stimulus candidates except for gender]
As another example, participants assigned to the SE condition within the high-justifiability/low-evaluability condition read the following:
In this study, we will ask you to rate the qualifications of two candidates from a foreign country for a paramedic job. For each candidate, we will give you multiple pieces of information, such as gender, earliest starting date, and score on MED Test (an exam for paramedics in the foreign country). Your goal is to give an overall rating for the candidate.
Please note:
(1) We ask different participants to use different pieces of information when giving their overall ratings. For you, we want you to give your rating by using only the following pieces of information, and ignoring the other pieces of information. In other words, your rating of the candidate should be only based on the following two pieces of information:
– Earliest starting date
– MED Test score
(2) *You may assign the same overall rating or different overall ratings to the two candidates.*
[Manipulation-check/comprehension questions; see Appendix C]
[Next page]
Now, please start to rate the candidates.

	Candidate H
Gender	*Female*
Highest education	*Bachelor*
Earliest starting date	*May 1st*
MED Test score	*1400*

How qualified is this candidate?
[Next page]
[Filler candidate, same as the first stimulus candidate except for gender]
[Next page]
[Filler studies]
[Next page]
In a previous study, we asked you to evaluate two candidates. Now, we ask you to evaluate two more.
[Same instructions as above]
Now, please rate the candidates.

	Candidate L
Gender	*Female*
Highest education	*Bachelor*
Earliest starting date	*May 1st*
MED Test score	*900*

How qualified is this candidate?
[Next page]
[Filler candidate, same as the second stimulus candidate except for gender]

7.2. Results and discussion

Here we report the results from all participants. Excluding participants who failed to answer all the manipulation-check/comprehension questions correctly did not qualitatively change the results. See Appendix C for details.

Although this experiment included four variables, we had no specific predictions for any 4-way interaction; rather, we were primarily interested in testing our proposition in the high-justifiability/low-evaluability and low-justifiability/high-evaluability conditions. In the following sections, we first report on evaluation mode × alternative 2-way interaction results in each of these conditions as well as in the other two attribute-type conditions. Then, we report on higher-order interactions to test the moderating effects of evaluability manipulation and justifiability manipulation on the relative effect of the attribute between JE and SE. For a summary of the results, see Table 2.

7.3. Results from each of the four attribute-type conditions

In the high-justifiability/low-evaluability condition, both our theory and existing research on evaluability predicted a JE amplification effect. To test the prediction, we conducted a 2 (evaluation mode: JE vs. SE, between-participants) × 2 (alternative: better score vs. worse score, within-participants) ANOVA. As expected, the analysis yielded a significant interaction effect, $F(1,130) = 24.81$, $p < .001$, indicating that the scores made a greater difference in JE ($M_{\text{better-score}} = 91.35$, $SD_{\text{better-score}} = 10.96$; $M_{\text{worse-score}} = 57.00$, $SD_{\text{worse-score}} = 18.52$, $F(1, 130) = 352.82$, $p < .001$) than in SE ($M_{\text{better-score}} = 79.21$, $SD_{\text{better-score}} = 16.09$; $M_{\text{worse-score}} = 58.26$, $SD_{\text{worse-score}} = 18.38$, $F(1, 130) = 112.75$, $p < .001$). The ANOVA also found a main effect of evaluation mode ($F(1,130) = 4.78$, $p = .03$) and a main effect of alternatives ($F(1,130) = 422.57$, $p < .001$).

In the low-justifiability/high-evaluability condition, our theory predicted an SE amplification effect. Supporting the prediction and replicating the results of Experiment 1B and Experiment 2, a 2 (evaluation mode) × 2 (alternative) ANOVA produced a significant interaction effect ($F(1, 133) = 6.11$, $p = .015$), indicating that the scores of the candidates made a greater difference in SE ($M_{\text{better-score}} = 75.41$, $SD_{\text{better-score}} = 22.83$; $M_{\text{worse-score}} = 61.23$, $SD_{\text{worse-score}} = 21.16$, $F(1, 133) = 30.90$, $p < .001$) than in JE ($M_{\text{better-score}} = 78.83$, $SD_{\text{better-score}} = 21.97$; $M_{\text{worse-score}} = 73.46$, $SD_{\text{worse-score}} = 26.30$, $F(1, 133) = 4.62$, $p = .033$). The ANOVA also found a main effect of evaluation mode ($F(1,133) = 4.81$, $p = .03$) and a main effect of alternatives ($F(1, 133) = 29.99$, $p < .001$).

In the high-justifiability/high-evaluability condition, a 2 (evaluation mode) × 2 (alternative) ANOVA found no interaction effect ($F(1,125) = 0.56$, $p = .454$), suggesting there was neither JE amplification nor SE amplification. The analysis only yielded a main effect of evaluation mode ($F(1,125) = 6.42$, $p = .013$) and a main effect of alternatives ($F(1,125) = 1366.33$, $p < .001$).

Likewise, in the low-justifiability/low-evaluability condition, a 2 (evaluation mode) × 2 (alternative) ANOVA found no interaction effect ($F(1,125) = 0.70$, $p = .405$), suggesting there was neither JE amplification nor SE amplification. The analysis found no main effect of alternatives ($F(1,125) = 1.31$, $p = .255$) but found a main effect of evaluation mode ($F(1,125) = 4.11$, $p = .045$).

7.4. Relationships between conditions

So far, we have reported the results of the four attribute-type conditions separately. Here, we report results combining multiple conditions. First, we conducted a 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) × 2 (justifiability: high vs. low) × 2 (evaluability: high vs. low) 4-way ANOVA on all the data, and we found no significant interaction effect: $F(1, 513) = 0.000$, $p = .995$.

Next, we compared one attribute-type condition with another by performing two 3-way ANOVAs, thereby testing the moderating effects of both the evaluability and justifiability manipulations.

To test the moderating effect of the evaluability manipulation, we conducted a 2 (evaluability: low vs. high) × 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) 3-way ANOVA. The ANOVA yielded a significant 3-way interaction effect, $F(1, 517) = 4.69$, $p = .031$, indicating that the evaluability manipulation moderated the attribute's relative effect between JE and SE. This result supported the

existing research on evaluability (e.g., Hsee & Zhang, 2010).

To test the moderating effect of the justifiability manipulation, we conducted a 2 (justifiability: low vs. high) × 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) 3-way ANOVA. The ANOVA also produced a significant 3-way interaction effect, $F(1, 517) = 5.42$, $p = .02$, indicating that the justifiability manipulation also moderated the attribute's relative effect between JE and SE. This result supported our proposition as well as the proposition by Bazerman et al. (1998) regarding the role of justifiability in JE versus SE decisions.

7.5. Discussion

This experiment tested the relative effects of four types of attributes—spanning all four combinations of justifiability and evaluability—between JE and SE. Instead of using different attributes in each case, we used the same attribute (test score) and manipulated its evaluability and justifiability. The findings of the experiment lend strong support to our theory and highlight the importance of both evaluability and justifiability in determining the relative effect of an attribute between JE and SE. Holding the justifiability of the attribute constant, the relative effect of the attribute between the two evaluation modes depends on its evaluability; holding the evaluability of the attribute constant, the relative effect of the attribute between the two evaluation modes depends on its justifiability.

8. General discussion

Attribute and evaluation mode are two basic aspects of decisions. While extensive extant research has studied the effect of evaluation mode on decisions, most of that research has examined alternatives that vary on two attributes simultaneously. The present research addresses a more fundamental question—whether a single attribute shows JE amplification or SE amplification, holding other attributes constant.

To address the question, we draw a distinction between two general characteristics—evaluability and justifiability—that underlie all specific attributes. We propose (a) that if an attribute is high in justifiability and low in evaluability, then the attribute will show JE amplification, and (b) that if an attribute is low in justifiability and high in evaluability, then the attribute will show SE amplification. While previous research on evaluability has tested and supported the first proposition, the present research is the first to test and find unequivocal evidence for the second proposition, which was originally inspired by the Bazerman et al. (1998) should-versus-want argument. By doing so, this research enriches our understanding of the relationship between the two basic aspects of decisions—attribute and evaluation mode.

This research helps reconcile two ostensibly competing accounts for the extensive JE-SE preference reversal findings documented in the literature. One account is the evaluability hypothesis (Hsee & Zhang, 2010; Hsee et al., 1999; Hsee, 1996a; Yeung & Soman, 2005), and the other is the should-want account proposed by Bazerman et al. (1998). Note that all JE-SE preference reversals explicitly or implicitly involve two attributes. The evaluability account ascribes JE-SE preference reversals to a difference in evaluability between the two attributes, while the should-want account ascribes JE-SE preference reversals to a difference in justifiability between the two attributes.

Most extant JE-SE preference-reversal findings can be classified in two categories. In one, the two attributes that underlie the preference reversal differ only in evaluability, and not in justifiability (e.g., González-Vallejo & Moran, 2001; Hsee, 1996a; Hsee et al., 2013; List, 2002). These preference reversals are consistent with the evaluability account and not with the should-want account. They are also consistent with Sher and McKenzie (2014) proposition that JE offers more information about the market distribution of the otherwise hard-to-evaluate attribute and therefore enables decision makers to evaluate it more effectively.

In the other category of JE-SE preference reversals, the constituent attributes differ not only in evaluability but also in justifiability (e.g., Bohnet et al., 2016; Shaffer & Arkes, 2009). Earlier in this article, we argued that the evaluability account is sufficient to explain this type of JE-SE preference reversal. Here, we emphasize that the should-want account is also sufficient to explain this type of JE-SE preference reversal. In order to show that the should-want account offers a unique explanation for a JE-SE preference reversal, the attributes underlying the preference reversal must vary only in justifiability and not in evaluability.

This research leaves a number of important questions unanswered. One such question is what makes an attribute unjustifiable. A decision maker may find an attribute unjustifiable for internal or private reasons, such as avoiding feelings of guilt (e.g., Amodio, Devine, & Harmon-Jones, 2007); for external or public reasons, such as following the rules of the law or avoiding criticism (Crandall & Eshleman, 2003; Crandall, Eshleman, & O'Brien, 2002); or for both internal and external reasons. This research does not distinguish between these reasons. However, identifying exactly why people perceive an attribute to be unjustifiable can improve our predictions of a differential effect between JE and SE. For example, if the reason is external, then the attribute will be more influential in SE than in JE only if the decision is made in public, but will be equally influential in JE and in SE if the decision is made in private. If the reason is internal, then the attribute will be more influential in SE than in JE regardless of whether the decision is made in public or in private.

Another type of low-justifiability factor, which we have not studied, is a superficial variable that may bias one's decision in SE, but is knowingly irrelevant. Like other low-justifiability attributes, we expect such attributes to exert less impact in JE than in SE. For example, in SE, people give higher estimates for $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ than for $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ (Tversky & Kahneman, 1974), but we predict that in JE, the difference would go away, because the order is clearly irrelevant. Similarly, Caruso, Gilbert, and Wilson (2008) found that in SE, people award more compensation to an accident victim if her pain is in the future than if it was in the past, but in JE, they award the same amount of compensation, because timing (future vs. past) is apparently irrelevant.

When testing our main proposition, we used alternatives that differed only on a low-justifiability attribute while holding everything else constant. What would happen if the alternatives also differed on some high-justifiability attributes in such a way that it is ambiguous which alternative is better? For example, suppose that two job candidates of different races differ on two high-justifiability attributes—education and work, with one candidate having a better education and the other having more work experience. In this case, even in JE, race may influence the recruiters' evaluations because the tradeoff between education and experience (the high-justifiability attributes) already makes the decision ambiguous (e.g., Hsee, 1996b; Norton et al., 2004). Would race (the low-justifiability attribute) exert additional influence in SE? It might, because SE affords even more freedom, but this is only our speculation.

This research tests only how the justifiability and evaluability of an attribute influence the relative effect of the attribute between JE and SE, but it does not discern whether the difference in effect between JE and SE is due to a change in weight (i.e., how much importance the decision maker attaches to the attribute) or a change in value differentiation (i.e., how much the decision maker is able to differentiate between the values of the alternatives). We posit that justifiability affects weight, and evaluability affects value differentiation. That is, the SE amplification effect happens because SE increases the weight of a low-justifiability attribute, and the JE amplification effect happens because JE increases the value differentiation of a low-evaluability attribute. Future research is needed to test this proposition empirically.

This research examines only the interaction between evaluation modes and alternatives, and not the main effect of evaluation mode. But

understanding potential JE-SE main effects is nevertheless important. For example, if jurors generally award harsher punishments in JE than in SE, then defendants may want to avoid JE regardless of the number and ethnicities of the victims. Some extant research has already explored JE-SE main effects (Hsee & Leclerc, 1998), but the research includes only high-justifiability attributes and may not apply to low-justifiability attributes.

This research also carries practical implications. Let us return to an example introduced earlier: Suppose you belong to a discriminated-against minority group, and you are applying for a job. You have the option to interview jointly with another candidate who is similar to you, except that he is not a minority, or you can interview alone. Intuitively, you may opt to interview alone, believing that SE would make your ethnic status less salient. However, our research recommends that you interview jointly with the other candidate, because ethnicity is already high in evaluability and already salient in SE, yet the low justifiability of ethnicity could inhibit the attribute's influence in JE.

In summary, this research distinguishes between justifiability and evaluability as two orthogonal determinants of whether a given attribute has a greater effect in joint evaluation or in single evaluation. By doing so, this research enriches our understanding of basic human decision processes, reconciles different interpretations of preference reversals in the literature, generates and tests new predictions, and suggests ways to improve decisions in applied contexts.

Acknowledgement

We thank Oleg Urminsky for developing the statistical analysis to test JE/SE differences, as described in Appendix B, and thank Alyssa Eldridge, Reid Hastie and Jiao Zhang for helpful comments on earlier drafts.

Appendix A. An illustration that a JE-SE preference reversal does not require either attribute to show SE amplification

Suppose that alternatives A and B involve a tradeoff between attribute X and attribute Y. Attribute X shows more JE amplification than attribute Y. Specifically, X creates 7 units of difference in JE and 3 units of difference in SE in favor of A, and Y creates 5.5 units of difference in JE and 4.5 units of difference in SE in favor of B. Assume that the overall preference for A over B in each evaluation mode is a linear combination of the differences of the two constituent attributes in that mode. Then, the overall preference for A over B in JE is $7 - 5.5 > 0$, and the overall preference for A over B in SE is $3 - 4.5 < 0$. In other words, the sign of the overall preference for A over B reverses between JE and SE. The example shows that even if neither attribute shows SE amplification, a JE-SE preference reversal could result.

Appendix B. A statistical test for a JE-SE preference reversal with mixed between- and within-participants manipulations

To test whether the effect of an attribute (e.g., ethnicity) differs between SE and JE is to test whether there exists a significant interaction effect between evaluation modes (JE vs. SE) and the alternatives that differ on the attribute (e.g., Belgian vs. Somalian ethnicity). However, in all but Experiment 2, the alternatives were presented between participants in SE and within participants in JE. This design makes ANOVA inapplicable. To test the significance of the interaction, we used the following contrast-code regression strategy, developed by Oleg Urminsky.

Let R_{JEA} and R_{JEB} denote participants' responses to the two alternatives (JEA and JEB) in the JE condition, and R_{SEA} and R_{SEB} denote participants' responses to the two alternatives in the two SE conditions (SEA and SEB). To test whether the effect of the underlying attribute differs between JE and SE is to test whether the following difference-in-difference quantity is significantly different from zero:

$$(R_{SEB} - R_{SEA}) - (R_{JEB} - R_{JEA})$$

To do this, we fit the following regression:

$$DV = \beta_1 \times Difference + \beta_2 \times SE + \beta_3 \times Difference \times SE$$

where β_1, β_2, and β_3 are estimated coefficients.

In the SE conditions, we set the following:

$DV = R_{SEA}$ in the SEA condition, or $= R_{SEB}$ in the SEB condition;
$SE = 1$;
Difference $= 0.5$ in the SEA condition, or $= 1.5$ in the SEB condition.

As a result, the regression for SE simplifies to:

$R_{SEA} = 0.5\,\beta_1 + \beta_2 + 0.5\,\beta_3$ (in the SEA condition)
$R_{SEB} = 1.5\,\beta_1 + \beta_2 + 1.5\,\beta_3$ (in the SEB condition)

This means that we can think of the difference between responses to the two alternatives in SE as $R_{SEB} - R_{SEA} = \beta_1 + \beta_3$

In the JE condition, we set the following:

$DV = R_{JEB} - R_{JEA}$;
$SE = 0$;
Difference $= 1$.

Since SE = 0, the last two terms in the regression drop out, and the regression for JE simplifies to:

$$R_{JEB} - R_{JEA} = \beta_1$$

Combining the regressions for SE and for JE, we can express the key difference-in-difference quantity as follows:

$$(R_{SEB} - R_{SEA}) - (R_{JEB} - R_{JEA}) = (\beta_1 + \beta_3) - (\beta_1) = \beta_3$$

Therefore, to test whether the difference-in-difference quantity differs from 0 is to test whether the parameter β_3 significantly differs from 0. If it does, then the effect of the attribute differs significantly between JE and SE.

Below are the steps for running the regression in SPSS. First, create four variables:

DV = $R_{JEB} - R_{JEA}$ in the JE condition, R_{SEA} in the SEA condition, or R_{SEB} in the SEB condition;
Difference = 1 in the JE condition, 0.5 in the SEA condition, or 1.5 in the SEB condition;
SE = 0 in the JE condition, or 1 in both the SEA and SEB conditions;
*Interaction = Difference * SE*.

Then run the following regression:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA = PIN(0.05) POUT(0.10)
/ORIGIN
/DEPENDENT DV
/METHOD = ENTER Difference SE Interaction.
```

The significance of the interaction coefficient indicates whether the effect of the attribute differs significantly between JE and SE.

Appendix C. Results of Experiment 3 based on participants who correctly answered all manipulation-check/comprehension questions

In all conditions, after reading the instructions but before evaluating the candidates, participants answered the following questions:

To make sure you understand our instructions, we want you to answer the following questions.
According to the instructions, should you consider each of the following factors when assigning your overall rating? (Yes or No)
- *Gender*
- *Highest education*
- *Earliest starting date*
- *MED Test score*

Suppose two candidates are identical on the factors you should consider, but different on the factors you should not consider. What should you do?
- *Give them the same overall rating*
- *Give them different overall ratings*

Suppose two candidates are different on the factors you should consider, but identical on the factors you should not consider. What should you do?
- *Give them the same overall rating*
- *Give them different overall ratings*

Of the original 521 participants, 447 answered all three questions correctly. The results below are based on these participants.

C.1. Results from each of the four attribute-type conditions

In the high-justifiability/low-evaluability condition, a 2 (evaluation mode: JE vs. SE, between-participants) × 2 (alternative: better score vs. worse score, within-participants) ANOVA yielded a significant interaction effect, $F(1,112) = 19.53$, $p < .001$, indicating that the scores of the candidates made a greater difference in JE ($M_{\text{better-score}} = 91.33$, $SD_{\text{better-score}} = 11.31$; $M_{\text{worse-score}} = 56.85$, $SD_{\text{worse-score}} = 18.52$, $F(1, 112) = 305.75$, $p < .001$) than in SE ($M_{\text{better-score}} = 77.72$, $SD_{\text{better-score}} = 16.33$; $M_{\text{worse-score}} = 56.02$, $SD_{\text{worse-score}} = 14.85$, $F(1, 112) = 105.23$, $p < .001$). The ANOVA also found a main effect of evaluation mode ($F(1,112) = 7.15$, $p = .009$) and a main effect of alternatives ($F(1,112) = 377.38$, $p < .001$).

In the low-justifiability/high-evaluability condition, a 2 (evaluation mode) × 2 (alternative) ANOVA produced a significant interaction effect ($F(1, 113) = 10.99$, $p = .001$), indicating that the scores of the candidates made a greater difference in SE ($M_{\text{better-score}} = 73.67$, $SD_{\text{better-score}} = 23.67$; $M_{\text{worse-score}} = 60.93$, $SD_{\text{worse-score}} = 20.74$, $F(1, 113) = 29.83$, $p < .001$) than in JE ($M_{\text{better-score}} = 77.84$, $SD_{\text{better-score}} = 22.72$; $M_{\text{worse-score}} = 76.09$, $SD_{\text{worse-score}} = 24.75$, $F(1, 113) = 0.56$, $p = .458$). The ANOVA also found a main effect of evaluation mode ($F(1,113) = 5.96$, $p = .016$) and a main effect of alternatives ($F(1, 113) = 19.14$, $p < .001$).

In the high-justifiability/high-evaluability condition, a 2 (evaluation mode) × 2 (alternative) ANOVA found no significant interaction effect ($F(1,106) = 3.24$, $p = .075$), suggesting there was neither JE amplification nor SE amplification. The analysis did not find a main effect of evaluation mode ($F(1,106) = 3.04$, $p = .084$), but found a main effect of alternatives ($F(1,106) = 1599.23$, $p < .001$).

Likewise, in the low-justifiability/low-evaluability condition, a 2 (evaluation mode) × 2 (alternative) ANOVA found no interaction effect ($F(1,108) = 1.32$, $p = .253$), suggesting there was neither JE amplification nor SE amplification. The analysis found no main effect of alternatives ($F(1,108) = 0.001$, $p = .980$), but found a main effect of evaluation mode ($F(1,108) = 4.62$, $p = .034$).

C.2. Relationships between conditions

So far, we have reported the results of the four attribute-type conditions separately. Here, we report results combining multiple conditions. First, we conducted a 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) × 2 (justifiability: high vs. low) × 2 (evaluability: high vs. low) 4-way ANOVA on all the data, and we found no significant interaction effect: $F(1, 439) = 1.45$, $p = .229$.

Next, we compared one attribute-type condition with another by performing two 3-way ANOVAs, thereby testing the moderating effects of the evaluability and justifiability manipulations.

A 2 (evaluability: low vs. high) × 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) 3-way ANOVA found a significant 3-way interaction effect, $F(1,443) = 3.99$, $p = .046$, which means that the evaluability manipulation moderated the attribute's relative effect between JE and SE. These results supported the existing theory on evaluability.

A 2 (justifiability: low vs. high) × 2 (evaluation mode: JE vs. SE) × 2 (alternatives: better score vs. worse score) 3-way ANOVA also found a significant 3-way interaction effect, $F(1,443) = 5.87$, $p = .016$, which means that the justifiability manipulation moderated the attribute's relative effect between JE and SE. These results supported our proposition regarding the role of justifiability in JE versus SE decisions.

References

Abrams, D., Bertrand, M., & Mullainathan, S. (2012). Do judges vary in their treatment of race? *Journal of Legal Studies, 41*(2), 347–383.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2007). A dynamic model of guilt implications for motivation and self-regulation in the context of prejudice. *Psychological Science, 18*(6), 524–530.

Bazerman, M. H., Tenbrunsel, A. E., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Making decisions with competing internal preferences. *Academy of Management Review, 23*(2), 225–241.

Bitterly, T. B., Mislavsky, R., Dai, H., & Milkman, K. L. (2014). Dueling with desire: A synthesis of past research on want/should conflict. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2403021.

Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science, 62*(5), 1225–1234.

Caruso, E. M., Gilbert, D. T., & Wilson, T. D. (2008). A wrinkle in time: Asymmetric valuation of past and future events. *Psychological Science, 19*(8), 796–801.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*(3), 414–446.

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology, 82*(3), 359–378.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44*(1), 20–33.

Dunn, T. L., Koehler, D. J., & Risko, E. F. (2017). Evaluating effort: Influences of evaluation mode on judgments of task-specific efforts. *Journal of Behavioral Decision Making, 30*, 869–888.

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*(2), 216–233.

Hodson, G., Dovidio, J. F., & Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin, 28*(4), 460–471.

Hsee, C. K. (1995). Elastic justification: How tempting but task-irrelevant factors influence decisions. *Organizational Behavior and Human Decision Processes, 62*(3), 330–337.

Hsee, C. K. (1996b). Elastic justification: How unjustifiable factors influence judgments. *Organizational Behavior and Human Decision Processes, 66*(1), 122–129.

Hsee, C. K. (1996a). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*(3), 247–257.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-valued options. *Journal of Behavioral Decision Making, 11*, 107–121.

Hsee, C. K., & Leclerc, F. (1998). Will products look more attractive when presented separately or together? *Journal of Consumer Research, 25*(2), 175–186.

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*(5), 576–590.

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology, 86*(5), 680–695.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5*(4), 343–355.

Hsee, C. K., Zhang, J., Wang, L., & Zhang, S. (2013). Magnitude, time, and risk differ similarly between joint and single evaluations. *Journal of Consumer Research, 40*(1), 172–184.

Krüger, T., Mata, A., & Ihmels, M. (2014). The Presenter's Paradox revisited: An evaluation mode account. *Journal of Consumer Research, 41*(4), 1127–1136.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255–275.

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge University Press.

List, J. A. (2002). Preference reversals of a different kind: The "more is less" phenomenon. *The American Economic Review, 92*(5), 1636–1643.

Mellers, B. A., Chang, S. J., Birnbaum, M. H., & Ordonez, L. D. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance, 18*(2), 347–361.

Montanari, K., Thomas B., Treyball, C., & Acikgoz, Y. (2017). Adding vs. averaging: How do job applicants evaluate job attributes? Working Paper, Appalachian State University.

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87*(6), 817–831.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115.

Nowlis, S. M., & Simonson, I. (1997). Attribute-task compatibility as a determinant of consumer preference reversals. *Journal of Marketing Research, 2*(34), 205–218.

Shaffer, V. A., & Arkes, H. R. (2009). Preference reversals in evaluations of cash versus non-cash incentives. *Journal of Economic Psychology, 30*(6), 859–872.

Sher, S., & McKenzie, C. R. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General, 143*(3), 1127–1143.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology, 37*(12), 2297–2306.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Yeung, C. W., & Soman, D. (2005). Attribute evaluability and the range effect. *Journal of Consumer Research, 32*(3), 363–369.