# Gamble Evaluation and Evoked Reference Sets:

# Why Adding a Small Loss to a Gamble Increases Its Attractiveness

Craig R. M. McKenzie
UC San Diego

Shlomi Sher
Pomona College

Draft of August 20, 2016

## Abstract

When presented with a gamble involving a chance of winning $9, participants rate it as only moderately attractive. However, when other participants are presented with a gamble that adds a chance of losing 5 cents -- resulting in gamble that is strictly worse -- they rate it as much more attractive. This surprising effect has previously been explained in terms of the small loss increasing the affective evaluability of $9. This paper argues for an alternative model, in which the baseline and small-loss gambles evoke different reference sets for comparison. In inferring a relevant reference set, people are sensitive to both the objective content and the framing of a gamble. The model distinguishes between two effects of evoked reference sets on behavior -- an obligatory (and normative) effect on scale interpretation, and an optional effect on subjective value. Five experiments provide strong evidence for the evoked reference set model. Data from attractiveness ratings suggest large and consistent reference set effects on scale interpretation, while data from choice tasks suggest that direct effects on subjective value may be less robust.

Context effects in decision making are said to occur when seemingly irrelevant changes to the choice environment affect judgments and choices (e.g., Huber, Payne, & Puto, 1982; Lichtenstein & Slovic, 1971; Payne, Bettman, & Johnson, 1993). Such effects are of interest in part because they appear to violate principles of rational choice. However, it is often difficult to know whether a change in context is irrelevant or instead provides information that a rational actor would utilize. Thus, it is crucial to analyze what information might be conveyed by a given change in context, as well as whether the information is sufficient to explain participants' behavior (e.g., Sher & McKenzie, 2006, 2011, 2014).

Slovic, Finucane, Peters, and MacGregor (2002) reported an intriguing effect in risky choice. Some participants were presented with a "standard" gamble with a 7/36 chance of winning $9 and a 29/36 chance of winning nothing. How attractive would it be to play the gamble one time? Participants' mean response was 9.4 on a scale ranging from 0 (not at all attractive) to 20 (very attractive). Other participants were presented with the same gamble, but a small loss component was added. Instead of a 29/36 chance of winning nothing, there was a 29/36 chance of losing 5 cents. Although the small-loss gamble is strictly worse than the standard gamble, the mean attractiveness rating increased to 14.9. Adding the small loss affected choices, as well. Compared to those presented with the standard gamble, participants presented with the small-loss gamble were more likely to prefer playing the gamble to receiving $2 for sure (61% vs. 33%).

To explain these seemingly counter-normative effects, Slovic et al. (2002) hypothesized that participants use an "affect heuristic", which assumes that, "in the process of making a judgment or decision, people consult or refer to the positive and negative feelings consciously or unconsciously associated with the mental representations of the task" (Bateman, Dent, Peters,

Slovic, & Starmer, 2007, p. 366). Slovic et al. proposed that, without the small loss outcome, it is difficult to know how good the $9 outcome is, since there is nothing to compare it with. Because it is difficult to evaluate $9 on its own, participants focus on the 7/36 probability, which is low, and therefore rate the gamble as only slightly attractive. By adding the small loss outcome, however, $9 becomes "evaluable" (e.g., Hsee, 1996) and "comes alive with feeling" (Slovic et al. 2002; Bateman et al., 2007), thereby increasing the gamble's appeal. We will refer to this explanation as the "affective evaluability" account.

## Attractiveness Ratings

While the affective evaluability account of the attractiveness results is plausible, it is not obvious that it would predict a large difference in ratings between the standard gamble and the small-loss gamble. That is, why does "lose 5 cents" make $9 come alive with much more feeling than "win nothing" does? Perhaps there is something special about a $0 outcome that hinders evaluability (e.g., it cannot enter into ratio calculations). However, Bateman et al. (2007) reported that a gamble with the second outcome described as "lose nothing" was rated as much more attractive than the same gamble with the second outcome described as "win nothing". The affective evaluability of the $9 outcome appears to be the same in the two $0-outcome gambles, so why are the resulting ratings so different? Bateman et al. suggested that the difference was due to the more "positive tone" of "lose nothing", but this explanation is different from the affective evaluability explanation (because evaluability is no longer a component). Parsimony is reduced when one account (affective evaluability) is used to explain the difference in ratings when the second outcome is "win 0" vs. "lose 5 cents" and a second account ("tone") is used to explain the difference when the outcomes are "win 0" vs. "lose 0". Finally, Bateman et al. showed that adding a small gain outcome of 5 cents also increased the attractiveness of the

standard gamble, but not as much as adding the small loss did. Why is the evaluability of $9 aided more by the small loss than by the small gain?

An alternative and more parsimonious explanation of the results is that the different gambles evoke different reference sets for comparison. When presented with a single gamble and asked how attractive it is, the natural question is: Compared to what? Indeed, this question must be answered by participants at some level in order to interpret the subjective attractiveness scale (e.g., what counts as a 12 on the scale?). Because the standard gamble involves only wins, we suspect that the evoked reference set tends to consist of other gambles involving only wins. While it may not be clear how appealing winning $9 is, winning nothing is the worst possible outcome when only gains are possible, thereby reducing the gamble's attractiveness. However, when the small loss is added, the resulting gamble includes both a winning and a losing outcome, and may evoke a reference set of gambles involving wins and losses. A 5-cent loss is about as good as a loss can be, thereby increasing this gamble's attractiveness relative to the reference set. This account also naturally explains why the gamble is rated as much more attractive when the second outcome is described as "lose 0" rather than "win 0", although all outcomes are the same: In a context of gambles involving wins and losses, losing $0 is the best possible loss, whereas in a context of gambles involving only wins, winning $0 is the worst possible outcome.

Note that, while the affective evaluability account focuses on the difficulty of putting the $9 outcome in context, the evoked reference set account focuses on how to put the entire gamble in context. We are suggesting that the small-loss effect on attractiveness ratings is due to the gambles being compared to the different reference sets that their outcomes or descriptions evoke.

Figure 1 illustrates the proposed account of how a single gamble is evaluated in isolation. To express an evaluation of a gamble on a given judgment scale, it is necessary to form some

internal representation of the value of the gamble, and also to construct an interpretation of the scale, in order to map the gamble's subjective value onto a specific scale level. We assume that a gamble is composed of not only its objective components (probabilities and outcomes), but also its description or framing (e.g., "win 0" vs. "lose 0"). The objective components of the gamble directly influence its subjective value, independent of the specific scale used to report evaluations. Highly probable large gains make a gamble valuable, for example.

When evaluations are expressed on a subjective scale (e.g., attractiveness), a reference set must be used to anchor the scale values. In Figure 1, the objective components of the gamble not only influence subjective value, but also influence which reference set is evoked. Imagine, for example, that the option being evaluated were a sure $100. Subjective value would likely be clear, but because the rating scale is bounded while monetary amounts are unbounded, the numerical rating provided by the participant must strongly depend on the range of payoffs under consideration. If the possible sure amounts were known to range between $0 and $100, then the $100 stimulus would presumably be mapped to the high end of the scale; but if the known range were $0 to $1,000, then $100 would fall near the low end. When only a single option is presented, the relevant range of outcomes is unspecified. What should participants do under such circumstances? One strategy would be for them to throw up their hands and respond either randomly or with the midpoint on the scale. At least some participants do not adopt this strategy, as different gambles lead to systematically different responses. Instead, participants apparently try to provide meaningful responses by answering the implicit "Compared to what?" reference set question as best they can. They appear to infer, perhaps implicitly, what the range of outcomes might be, even if only in rough and uncertain terms. It is plausible that the presentation of a gamble involving only wins is likely to evoke a reference set of other gambles

involving only wins. By contrast, a gamble with potential win and loss outcomes is more likely to evoke a reference set of gambles involving both wins and losses. That is, participants may naturally evoke the most parsimonious reference set, at least in qualitative terms.

This pattern of behavior is reasonable if participants assume that the focal gamble has been drawn from a larger relevant consideration set. For example, a gamble involving both wins and losses could not have been sampled from a population of gambles involving only wins. A single gamble can provide compelling evidence regarding the population from which it was sampled. Furthermore, human decision makers are highly sensitive to such evidence, drawing strong population inferences from small samples when little other information is available (Sher & McKenzie, 2014; see also Vul, Goodman, Griffiths, & Tenenbaum, 2014).

Importantly, how the gamble is described (or framed) can also influence the evoked reference set. A gamble with outcomes of "win $9" and "win $0" and a gamble with outcomes of "win $9" and "lose $0" have the same objective outcomes, but different descriptions. The former description may be more likely to evoke a reference set of gambles involving only wins, and the latter a reference set involving wins and losses. Whether this is reasonable depends on whether the description provides evidence about the reference set (context) from which the single gamble was sampled. If “speakers” are more likely to describe a $0 outcome as “win nothing” when other gambles in the immediate context involve only wins compared to when the other gambles involve wins and losses, then it would be reasonable for a “listener” to infer that other gambles in the set involve only wins when the $0 outcome is framed as “win nothing” (see McKenzie & Nelson, 2003; Sher & McKenzie, 2006). Experiment 3 provides evidence that people are, in fact, more likely to describe a $0 outcome as “win nothing” (rather than “lose

nothing") when other gambles in the set involve primarily wins rather than a mix of wins and losses.

The context evoked by the single gamble enables the decision maker to interpret the subjective rating scale. According to the model in Figure 1, subjective value is then expressed via the rating scale, with gambles whose value is typical for the evoked reference set receiving ratings near the midpoint. Therefore, when the reference set is diminished, any fixed level of subjective value maps onto a higher numerical rating. Assume that the standard and small-loss gambles have roughly equal subjective value (e.g., their expected values differ by only 4 cents). Because of the different evoked reference sets, the gambles would nonetheless be rated differently. Compared to the standard gamble, the small-loss gamble would fare better against its (inferior) evoked reference set and would therefore receive a higher attractiveness rating. Similarly, compared to the gamble with "win nothing" as its second component, the gamble with "lose nothing" as its second component would fare better against its (inferior) evoked reference set and would receive a higher rating.

In a theoretical article, Kahneman and Miller (1986) proposed a similar explanation of the (then unpublished) small-loss effect on attractiveness ratings. They suggested that the standard gamble and the small-loss gamble evoke different norms (an evoked set of exemplars that serves as a representation of a normal outcome or category member, roughly analogous to a reference set) because the gambles differ in terms of "the presence or absence of risk of loss" (p. 142). "The [small-loss] bet appears very favorable among bets that involve a risk of loss, but a modest chance to win $9 is mediocre in a context of purely positive prospects" (pp. 141-142). Our account can be viewed as both a refinement and an extension of their explanation: We distinguish between mechanisms whereby reference sets can affect scale interpretation versus

subjective value (see below), relate the effects of gamble descriptions on evoked reference sets to the effects of known reference sets on descriptions, and, for the first time, present empirical evidence in support of the proposed explanation.

The inclusion of positive and/or negative outcomes can influence reference sets in a variety of settings. For example, McGraw, Larsen, Kahneman, and Schkade (2010) noted that loss aversion ("losses loom larger than gains") is robust in choice tasks but often disappears when participants rate how gains and losses would make them feel. They argued that, while choice compels participants to directly compare gains and losses, asking for ratings often leads participants to evaluate the outcome relative to other outcomes with the same valence. That is, when asked to rate the intensity of losing $15, it is natural to rate it relative to other losses, and when asked to rate a gain of $15, it is natural to compare it with other gains. Because the psychological scale for gains may differ from that for losses, comparing such ratings can be misleading. In another study, following up on an experiment by Birnbaum (1999), Leong, McKenzie, Sher, and Müller-Trede (2016) elicited subjective ratings of the "largeness" of the numbers 2 and -2 in a between-subjects design. On average, participants rated the number 2 as smaller than the number -2, presumably because 2 evokes a reference set of natural numbers (relative to which 2 is small) while -2 evokes a reference set of integers (relative to which -2 is neither large nor small).[1]

The preceding description of the rating process is essentially a simple normative analysis and suggests that the small-loss effect on attractiveness ratings is not irrational. The effect can be explained by the obligatory use of an implicit context to interpret an ambiguous judgment scale. However, in Figure 1, the broken arrow from "evoked context" to "subjective value" represents another potential route in the model that, though neither obligatory nor normative, is

psychologically plausible. Evoked contexts might influence ratings not just via scale interpretation, but also via effects on subjective value. It is possible, for example, that adding the small loss not only increases ratings because an inferior reference set is used to interpret the scale, but also because the inferior reference set makes the gamble "feel" more valuable by subjective contrast.

How can we know whether attractiveness ratings are being driven by scale interpretation only, or by both scale interpretation and changes in subjective value? One way is to eliminate scale interpretation as a feasible route and see if the small-loss effect persists. If it does, then this would indicate that evoked context is influencing subjective value over and above its effect on scale interpretation. This possibility is most naturally examined by using choices rather than subjective ratings as the dependent measure.

## Choices

Recall that Slovic et al. reported that, when given a choice between playing the gamble one time and receiving $2 for sure, participants more often preferred the gamble when the 5-cent loss was added. That is, participants not only rated the small-loss gamble as more attractive, but also more often preferred it to a sure amount. Because there is no subjective scale to interpret when making a choice, no reference set needs, in principle, to be evoked. Instead, the two options, the gamble and a certain $2, only need to be compared to each other. Whether the gamble was drawn from a set involving only wins or from a set involving wins and losses has no obvious normative relevance for preference between the gamble and a sure gain of $2.

Choice effects can nonetheless be explained by the model in Figure 1 if reference sets (a) are evoked even in choice tasks and (b) influence subjective value (broken arrow). Specifically, the subjective value of the gamble may exhibit a contrast effect relative to the evoked reference

set. For example, by adding the small loss to the standard gamble in the choice task, the subjective value of the gamble might be enhanced by an evoked reference set of other gambles involving wins and losses, thereby making it more likely to be selected over the $2.

Because choice tasks eliminate the need for scale interpretation, the model in Figure 1 implies that, if there is any effect on choice of adding the small loss to the standard gamble, then an evoked reference set is influencing subjective value. Whereas there are two routes in the model for evoked reference sets to influence subjective ratings, there is only one route to influence choices. By examining the pattern of results in judgment vs. choice tasks, we should be able to draw conclusions about the relative influence of evoked reference sets on subjective value in both tasks. For example, if the small-loss effect is much weaker, or non-existent, for choices as compared to ratings, this would suggest not only that evoked reference sets have little or no influence on choice, but also that scale interpretation is accounting for most or all of the effect in the ratings task. However, if the small-loss effect is similar for both ratings and choices, this would suggest that evoked reference sets are playing a large role in both tasks, and that scale interpretation is playing a relatively small role in ratings.

Bateman et al. (2007) and Slovic et al. (2002) reported a total of four choice experiments examining the small-loss effect. They looked at participants' country of origin (US and UK) and the size of the sure amount that was offered along with the gamble (2 dollars/pounds and 4 dollars/pounds). The two experiments using 2 dollars/pounds as the sure amount resulted in a significant effect: Participants preferred the gamble to the sure amount more often when the small loss was added to the standard gamble. The two experiments using 4 dollars/pounds did not result in a significant effect. It might be that the larger sure amount resulted in a floor effect and masked the small-loss effect. However, it could also be that the effect on choices is less

robust than the effect on ratings. More data are needed to see if the small-loss effect on choice is as strong and consistent as the effect on ratings.

## Overview of Experiments

In what follows, we report five experiments that replicate the small-loss effect on ratings, test competing predictions of the affective evaluability and evoked reference set accounts, and address the two routes, depicted in Figure 1, whereby evoked contexts can influence behavior. Experiment 1 replicates and extends earlier research by having participants provide attractiveness ratings and make choices when the second outcome for the "win $9" bet is either "win 0", "lose 5 cents", or "lose 0". Experiment 2 provides a critical test of the competing accounts by not only manipulating the small outcome, but also manipulating whether the large outcome is winning $9 or losing $9. The affective evaluability account predicts that adding the 5-cent loss will make the $9 loss "come alive with feeling" and therefore be less attractive, whereas the evoked reference set account predicts that the 5-cent loss will evoke an inferior reference set and thus make the gamble more attractive. Experiment 3 tests whether the reference sets that different gambles appear to evoke are reasonable. We do so by reversing the process: Rather than manipulate the gambles and infer (via attractiveness ratings) which reference set is evoked, we manipulate the reference set -- whether a set of gambles involves only gains or both gains and losses -- and see if this influences whether a gamble outcome of $0 is described as "win nothing" or "lose nothing". Finally, Experiments 4 and 5 take a closer look at whether adding a small loss influences choices, not just attractiveness ratings.

## Experiment 1

Our first experiment involves three conditions that have not been tested together previously: The standard gamble (henceforth W9/W0), the small-loss gamble (W9/L0.05), and

the “lose nothing” gamble (W9/L0). The first two conditions have been compared (Bateman et al., 2007; Slovic et al., 2002), as have the first and third (Bateman et al., 2007), but the three conditions have not been examined within a single study. The evoked reference set account makes a clear prediction about the ordering of attractiveness ratings in the three conditions: W9/W0 < W9/L0.05 < W9/L0. Recall that Bateman et al. provided different explanations of the difference in results between the first two conditions (affective evaluability) and between the first and third conditions (tone), whereas the evoked reference set account can parsimoniously explain the differences between all three.

In addition, choice data have not previously been collected for the W9/L0 gamble, and we do so in this experiment. Comparing choices (between each gamble and a sure amount) across these three gambles is expected to shed light on any effect of evoked reference sets on subjective value.

Relatedly, we also asked some participants to provide willingness-to-pay (WTP) judgments (i.e., the most they would be willing to pay to play the gamble one time). WTP judgments have not previously been used for these gambles, and they are potentially useful because, unlike choices, they provide a scaled response while, unlike attractiveness ratings, they do not require participants to impute contextual meaning to an under-specified subjective scale.

Finally, previous experiments comparing responses between the W9/W0 and the W9/L0.05 gambles have confounded the gamble with the manner in which the second outcome is communicated to participants (Bateman et al., 2007; Slovic et al., 2002). The W9/L0.05 gamble has always been presented explicitly as:

7/36 chance of $9

29/36 chance of losing 5 cents

By contrast, when participants have been presented with the W9/W0 gamble, the 29/36 chance of winning nothing is not placed right below the first outcome, but is instead stated in the general instructions to participants (see Bateman et al.'s Figure 1). Thus, one reason for the small-loss effect, as mentioned by Bateman et al. (see their Footnote 6), could simply be that affective evaluability for the W9/L0.05 gamble is higher because attention is drawn to the small-loss outcome by placing it next to the $9 outcome. To test this possibility, the current experiment made the implicit "win nothing" outcome in Slovic et al. and Bateman et al. explicit by placing "29/36 chance to win $0.00" immediately below "7/36 chance to win $9.00".

Method

Participants were 308 UC San Diego undergraduates who received partial course credit (65% female, mean age = 20). They were randomly assigned to one of three conditions: W9/W0, W9/L0.05, or W9/L0. The task was a paper-and-pencil survey that was part of a series of experiments taking less than an hour. After a general introductory page, participants in the W9/W0 condition read the following:

> Imagine that you have the opportunity to play the gamble below one time for real money. The outcome is determined by spinning a wheel of fortune with 36 areas of equal size. Seven of the areas are green, and 29 of the areas are red. If the spinner lands on a green area, you win $9.00. If the spinner lands on a red area, you win $0.00.
>
> So the gamble is this:
>
> 7/36 chance to win $9.00
>
> 29/36 chance to win $0.00

In the W9/L0.05 and the W9/L0 conditions, "win $0.00" was replaced by "lose $0.05" and "lose $0.00", respectively. All participants reported how attractive they found the gamble by circling

one number on a scale ranging from 1 (not at all attractive) to 20 (very attractive). Half of the participants in each condition then reported whether they would prefer to play the gamble once or receive $2.00 for sure, and the other half reported "willingness to pay" (WTP) by answering the question, "What is the most you would be willing to pay to play this gamble one time?".

Results and discussion

Attractiveness ratings. Figure 2 shows the results for the attractiveness ratings (N = 308). As predicted by the evoked reference set account, mean ratings conformed to the order W9/W0 < W9/L0.05 < W9/L0 (Ms = 8.1, 10.9, 14.1). A one-way ANOVA revealed an effect of Gamble ($F(2, 305) = 26.2$, $p < .001$), and pair-wise contrasts showed that each condition mean was different from the other ($t$s > 3.5, $p$s < .001). Because participants found the W9/W0 gamble to be less attractive than the W9/L0.05 gamble, even though the "win $0.00" outcome was explicit, the original effect is not due to the fact that "win $0.00" was implicit. Nor is it due to the fact that there is something special about a $0 outcome that hinders affective evaluability, because W9/L0 was considered more attractive than W9/L0.05. Moreover, affective evaluability is presumably held constant for the W9/W0 and W9/L0 gambles, and the latter was judged much more attractive than the former.

Choices. Figure 3 shows how often participants chose to play each gamble rather than receive $2 for sure (n = 153), and the pattern is similar to that for attractiveness ratings. Participants in the W9/W0 condition chose the gamble least often and those in the W9/L0 condition chose it most often (Ms = .33, .45, .59). A log-linear analysis showed an effect of Gamble, $\chi^2 = 6.8$, $p = .034$. Separate chi-squared tests, though, revealed that only the W9/W0 and W9/L0 conditions were significantly different from each other, $\chi^2(1, N = 102) = 6.7$, $p = .017$. (For W9/W0 vs. W9/L0.05, $p = .23$, and for W9/L0.05 vs. W9/L0, $p = .31$.) Affective

evaluability alone cannot explain the difference between the two $0-outcome gambles. However, the pattern of results is consistent with the evoked reference set account, assuming that reference sets are affecting subjective value.

Willingness to pay. The WTP results were difficult to interpret (n = 155). Indeed, 24% of participants reported being willing to pay more than $9 -- which is the most they could win by playing the gamble. After eliminating these participants, mean WTP was $2.59, $3.11, and $2.54 for the W9/W0, W9/L0.05, and W9/L0 conditions, respectively. A one-way ANOVA revealed no effect of gamble, $F(2,115) < 1$. One way to interpret this lack of effect is that evoked reference sets affect the construal of ambiguous subjective scales, but that they do not affect the use of a fully interpreted scale like WTP. However, the need to eliminate 24% of responses raises obvious concerns about the meaningfulness of these data. For reasons that are unclear to us, participants apparently had difficulty understanding the WTP question, and we did not include it in subsequent experiments.

In sum, Experiment 1 confirmed the predictions regarding attractiveness ratings and choices made by the evoked reference set account, and cannot be fully explained by affective evaluability. The effects of gamble on choice suggest that part of the effect on attractiveness ratings may be due to inflated subjective value, not just scale interpretation. The effects on choice appear less robust than those on ratings, though, which suggests that scale interpretation may be doing much of the work in the case of attractiveness ratings. However, the sample size for the choice data was only half that for the ratings data, thereby limiting power.

## Experiment 2

Experiment 2 was designed to provide an even more direct test of the evoked reference set and affective evaluability accounts. Two new conditions were created that were identical to

W9/W0 and W9/L0.05, but instead of a 7/36 chance to win $9, there was a 7/36 chance to lose $9. The affective evaluability account predicts that adding the small loss will make the $9 loss more evaluable and therefore attractiveness should decrease. By contrast, the evoked reference set account predicts that the L9/W0 gamble evokes a reference set of gambles involving wins and losses, winning $0 is bad in such a context, and hence attractiveness should be low. When the small loss is added (L9/L0.05), the gamble only involves losses. A reference set consisting of gambles involving only losses is evoked, the small loss is good in this context, and attractiveness should increase.

In addition, participants were again asked to choose between one gamble and a sure amount to shed additional light on whether inflated subjective value is influencing choices and, by extension, ratings.

Method

Participants were 220 UCSD undergraduates who received partial course credit (73% female, mean age = 20). They were randomly assigned to one of four conditions (n = 55 in each), and the experiment took the form of a paper-and-pencil survey. In two conditions, participants were presented with either the standard gamble (W9/W0) or the small loss gamble (W9/L0.05) as in Experiment 1. In the other two conditions, the second outcomes were the same as the first two conditions, but the first outcome involved a 7/36 chance to lose $9 (L9/W0 and L9/L0.05, respectively). In Experiment 2, as in previous published studies (Bateman et al., 2007; Slovic et al., 2002), the "win nothing" outcome was not explicit (i.e., it was not placed below "7/36 chance to win [lose] $9.00"), but was stated in instructions accompanying the gamble. The stimuli were virtually identical to those presented in Bateman et al.'s Figure 1 (including a visual depiction of a wheel of fortune). Because the new gambles involved only losses, we changed the

attractiveness scale to range from -10 (extremely unattractive) to +10 (extremely attractive). Participants in the two W9 conditions then reported whether they would prefer to play the gamble once or to receive $2 for sure, and those in the two L9 conditions reported whether they would prefer to play the gamble once or to lose $2 for sure.

Results and discussion

Figure 4 illustrates the attractiveness results. The left two bars show that the usual effect was replicated: Adding the small loss to the W9/W0 gamble made it more attractive, $t(108) = 3.31$, $p = .001$ (Ms = 0.6, 3.7). The right two bars represent the critical comparison, and show that adding the small loss to the L9/W0 gamble made it more attractive, as well, $t(107) = 3.55$, $p < .001$ (Ms = -6.5, -2.4). This is consistent with the evoked reference set prediction and is opposite the affective evaluability prediction.

Figure 5 shows how often participants preferred to play the gamble (rather than receive $2 for sure in the W9 conditions or lose $2 for sure in the L9 conditions). Adding the small loss increased choices slightly in both the W9 (.44 to .49) and the L9 cases (.71 to .76), but the increases were not close to significant ($\chi^2$s < 1, ps > .5). This is inconsistent with affective evaluability, which predicts that the 5-cent loss will increase choices for the W9/W0 gamble and decrease choices for the L9/W0 gamble. The choice results suggest that evoked reference sets did not have much influence on subjective value (broken arrow in Figure 1). Nonetheless, the results for attractiveness ratings suggest that evoked reference sets exert a consistently strong influence on scale interpretation.

These results also speak to whether adding the small loss affects how participants interpret the $9 outcome (as predicted by the affective evaluability account) or interpret the gamble more generally (as predicted by the evoked reference set account). On the former

account, the effects would have gone in opposite directions for the $9 gain and the $9 loss gambles. The fact that adding a small loss increased ratings for both gambles indicates that the evaluability of the $9 outcome is not central to the effect.

## Experiment 3

Thus far we have tested the competing accounts in part by manipulating gamble descriptions and seeing how this affects which reference sets are apparently evoked (as revealed by attractiveness ratings). In Experiment 3, we reversed the process: We manipulated a gamble's reference set to see how this affects the gamble's description. The evoked reference set account assumes that gamble descriptions evoke certain reference sets because the descriptions are a signal from a speaker (the experimenter, in this case) about the reference set. For example, it assumes that speakers are more likely to describe a $0 outcome as "win $0" (rather than "lose $0") when the gambles in the reference set consist of wins, compared to when the gambles in the set consist of both wins and losses. We noted earlier that the objective components of a gamble provide information about the reference set from which it was selected (e.g., a gamble involving wins and losses could not have been sampled from a set of gambles involving only wins). Adapting a frame selection task used in earlier research on "information leakage" (Sher & McKenzie, 2006), the current experiment tests the prediction that how a gamble is framed also provides information about the relevant reference set. If the prediction is confirmed, it would indicate that the reference sets that different gamble descriptions evoke are not only predictable, but reasonable.

### Method

Participants were 110 UCSD undergraduate students who received partial course credit (66% female, mean age = 20). They read over a table of 12 gambles with the knowledge that

one of the gambles would be selected and they would have to describe that gamble to a friend, who would then decide whether to play the gamble. (For details, see the Appendix.) The gambles in the table were composed of two possible outcomes and their probabilities (e.g., Gamble 10 was a 7/36 chance of $9 and a 29/36 chance of $0). The first outcome was always a gain (i.e., a positive dollar amount). Participants randomly assigned to the Win/Win condition (n = 55) saw gambles whose second outcome was almost always a gain (only 2 of the 12 values were negative), whereas those in the Win/Lose condition (n = 55) saw gambles whose second outcome was never a gain. In the latter case, 11 of the 12 second outcomes were negative, and one (Gamble 10) was $0. The two tables were identical except that all the positive second outcomes in the Win/Win table were made negative by placing a negative (-) sign in front of each to create the Win/Lose table. Gamble 10, however, was identical in both tables.

After reading the table, participants were asked to describe Gamble 10 to a friend (see the Appendix). They had to fill in the probability, dollar value, and valence (win vs. lose) for both outcomes. The key dependent measure was how often participants described the second outcome as a chance to "win" $0 vs. "lose" $0.

Results and discussion

Three participants were excluded from the analysis because they did not fully complete the questionnaire. In the Win/Win condition, 64% of participants described the $0 outcome as "win $0", whereas in the Win/Lose condition, only 24% did so, $\chi^2(1, N = 107) = 17.44$, $p < .001$. As expected, in a context primarily involving gains, the $0 outcome was more likely to be described as winning nothing compared to a context involving gains and losses. This indicates that describing the second outcome as "win nothing" -- as the standard gamble does -- should signal to the listener that the relevant reference set of gambles primarily involves gains (rather

than gains and losses). Similarly, describing the outcome as "lose nothing" should signal that the reference set of gambles involves a mixture of gains and losses (rather than only gains).

## Experiment 4

The attractiveness results of Experiments 1 and 2 support the evoked reference set account of why seemingly inconsequential changes to gambles have large effects. A single account can explain both why adding a small loss to a gamble increases its rated attractiveness, as well as why there is a large difference in response to gamble descriptions that are logically equivalent – i.e., the W9/W0 and W9/L0 gambles. Furthermore, the evoked reference set account raises questions about the counter-normative interpretation of differences in rated attractiveness, because attractiveness ratings require a reference set, and information about the gamble's reference set is provided by both the gamble's objective components and by how those components are framed (Experiment 3).

The choice results also clearly support the evoked reference set account over affective evaluability, but they were less clear with respect to whether evoked reference sets were influencing choices. Because no scale interpretation is required for the choice task, any influence of references sets on choice suggests that the effect is driven via subjective value (the broken arrow in Figure 1). This is not only difficult to justify normatively, but it would also suggest that this path in the model may be influencing attractiveness ratings, over and above the (necessary) effect of scale interpretation. While the effect on attractiveness ratings in Experiments 1-2 were large and consistent, the choice data were equivocal. The results of Experiment 1 seemed to indicate that evoked reference sets were influencing choice (Figure 3), but only one of the three pairwise comparisons was significant (W9/W0 vs. W9/L0). In Experiment 2, adding the small loss had no significant effect on choices for either the W9/W0

gamble or the L9/W0 gamble, though the small effects were in the direction expected if reference sets were influencing subjective value (Figure 5). In the current experiment, we examined choices only and used a larger sample in order to see if adding a small loss affects choices.

Method

The participants were 249 UCSD undergraduate students who received partial course credit for participating (55% female, mean age = 20). They were randomly assigned to either the W9/W0, W9/L0.05, or W9/L0 gamble (presented as in Experiment 1) and asked whether they would prefer to play the gamble or receive $2 for sure; there was no attractiveness ratings task. The experiment was a paper-and-pencil survey.

Results and discussion

Figure 6 shows how often participants chose each gamble over $2. Although W9/L0 resulted in the largest number of gamble choices, W9/W0 and W9/L0.05 were identical (Ms = .73, .63, .63). A log-linear analysis revealed no effect of gamble on choices, $\chi^2 = 2.7$, $p = .26$.

Note that none of our three experiments involving choices has revealed a significant effect on choice of adding the 5-cent loss to the standard gamble. According to the model in Figure 1, the results thus far suggest that the effect of adding the small loss on ratings is largely because evoked reference sets are influencing scale interpretation, not inflating subjective value.

## Experiment 5

The size of the effect of gamble on choice varied across the three choice experiments reported above. One difference between Experiment 1 (where we found the largest effect) and Experiment 4 (the smallest effect) is that the former asked for attractiveness ratings before making a choice, while the latter asked for only a choice. Because the attractiveness ratings task requires the active use of a reference set, it is possible that this initial effect of reference set

carries over to a subsequent choice task. When only a choice is required, however, reference sets may be less often evoked and/or receive less attention, and therefore have less of an effect. This would still not explain why the effect of adding the small loss to the standard gamble was so small in Experiment 2 (where attractiveness ratings always came first), but examining the effect of ratings preceding choice may clarify the relation between evoked reference sets and choices. To this end, participants in Experiment 5 were presented with either the W9/W0, W9/L0.05, or W9/L0 gamble and provided an attractiveness rating and made a choice. However, we manipulated which response came first. Of interest is whether there is a larger effect of the gamble on choice when choice comes after, rather than before, the attractiveness rating.

Method

The participants were 496 UCSD undergraduates who received partial course credit for participation (73% female, mean age = 20). They were randomly assigned one of the three gambles, with half of the participants answering the attractiveness question first, and half making a choice first. Also manipulated was whether the choice question listed the gamble or the sure amount first.

A few changes were made to this experiment. First, the choice was between playing the gamble one time and receiving $3 (rather than $2) for sure. The sure amount was increased because preference for the gamble exceeded 70% in some conditions in the earlier experiments, and we wanted to avoid any ceiling effects. Second, the experiment was run on computer to ensure that participants, when answering the first question, could not look ahead to see the second question and, after answering the second question, could not change their answer to the first question. Finally, in order to minimize response input errors, the attractiveness scale consisted of 1-digit numbers, ranging from 1 (not at all attractive) to 9 (very attractive).

Results and discussion

The top panel of Figure 7 shows attractiveness ratings as a function of gamble and whether the attractiveness rating or choice came first. A 3 (Gamble: W9/W0, W9/L0.05, W9/L0) x 2 (Question Order: rating first, choice first) between-participants ANOVA on attractiveness ratings revealed a main effect of Gamble ($F(2,490) = 38.6$, $p < .001$), with the usual order emerging: W9/W0 < W9/L0.05 < W9/L0. There was also a main effect of Question Order ($F(1,490) = 13.1$, $p < .001$); ratings were higher when they came first rather than second ($M$s = 5.7 vs. 5.0). Although the pattern of ratings across the three gambles appears somewhat flatter after choice compared to before choice, the interaction was not significant, $p = .15$.

Looking only at attractiveness ratings when they came first (as in Experiments 1 and 2; black bars in the top panel), all pairwise contrasts were significant ($t$s > 2.8, $p$s < .01). However, when choice came first (gray bars), W9/W0 was different from both W9/L0.05 and W9/L0 ($t$s > 5.4, $p$s < .001), but the difference between W9/L0 and W9/L0.05 disappeared ($t < 1$).

The bottom panel shows the results for choices. A 3 x 2 log-linear analysis on choices revealed only an effect of Gamble, $\chi^2 = 21.9$, $p < .001$. When ratings came first, there was the usual pattern: W9/W0 < W9/L0.05 < W9/L0. The difference between W9/W0 and W9/L0.05 was marginally significant ($\chi^2(1, N = 163) = 3.2$, $p = .074$), but the other contrasts were significant ($\chi^2$s > 4, $p$s < .05). When choice came first, W9/W0 was different from the other two gambles ($\chi^2$s > 4.8, $p$s < .03), but the difference between W9/L0 and W9/L0.05 disappeared ($p$ > .5).

The pattern of results was consistent across response mode, but which response mode came first had a small effect. The main change was that the difference between the W9/L0 and W9/L0.05 conditions effectively vanished for both choice and attractiveness ratings when choice

was first (though neither interaction was significant). The difference between W9/W0 and the other two gambles remained, however. Interestingly, the difference between choices for the W9/W0 and W9/L0.05 gambles was significant for the first time in our four experiments examining choices (thereby replicating Slovic et al., 2002), and occurred when choice came first.

Generally, it appears that there is an effect of gamble on choices, especially when ratings come first. The pattern of results is consistent with the evoked reference set account that assumes reference sets influence choice. (Recall that affective evaluability predicts no difference between the W9/W0 and W9/L0 conditions.) This suggests that subjective value is being influenced by evoked reference sets. However, in contrast to the effect on attractiveness ratings, the effect of gamble on choices has been variable, and often statistically non-significant, across the four experiments examining choices. Thus, there does seem to be some influence of evoked reference sets on choice, but the influence is relatively small and inconsistent.

## General Discussion

The present studies build on an intriguing finding reported by Slovic et al. (2002), in which adding a small loss to a gamble increases its judged attractiveness. Since the small-loss gamble is strictly worse than the standard gamble, these between-subjects results are puzzling. Slovic et al. and Bateman et al. (2007) appealed to the affect heuristic to explain the finding. Their idea was that it is difficult to know how good $9 is, and adding the small loss makes the $9 outcome “come alive with feeling”. The “win $0” outcome was said to hinder the affective evaluability of $9. However, Bateman et al. also presented a gamble to participants with the second outcome described as “lose nothing” rather than “win nothing”, and ratings were even higher. Because affective evaluability is presumably held constant for these two $0-outcome gambles, that construct cannot explain this effect. Bateman et al. instead suggested that the more

positive tone of “lose nothing” compared to “win nothing” explained the difference. But it seems unparsimonious to require different explanations to account for differences in ratings when the second outcomes are “win 0” vs. “lose 5 cents” (affective evaluability) and when they are “win 0” vs. “lose 0” (tone).

We proposed that the pattern of results can be explained more parsimoniously by positing that different outcome valences (win vs. lose) evoke different reference sets for comparison. When evaluating a single gamble in isolation on the subjective attractiveness scale, participants must answer the implicit question, “How attractive compared to what?”. Because the standard gamble (W9/W0) involves only wins, participants tend to compare it to other gambles involving only wins. Relative to that reference set, the gamble does not fare well because winning $0 is the worst possible outcome. However, adding a small loss (W9/L0.05) or reframing the neutral outcome as a zero loss (W9/L0) results in a different reference set for comparison, namely, gambles involving wins and losses. Losing 5 cents (or nothing) is almost the best (or the best) possible loss, so the gamble’s relative attractiveness increases. Because different reference sets are evoked by the different outcome valences, a dominated gamble can be rated as more attractive in a between-subjects design. The evoked reference set account predicts a specific order of attractiveness judgments (W9/W0 < W9/L0.05 < W9/L0) that was confirmed in Experiment 1.

Experiment 2 added two new conditions to examine the effect of adding a small loss to the risk of a $9 loss. The affective evaluability account predicts that the small loss should make the $9 loss more evaluable and therefore be rated as less attractive. By contrast, the evoked reference set account predicts that, because the new small-loss gamble only involves losses, and 5 cents is about as good as a loss can be, it will be rated more attractive than the L9/W0 gamble,

which is a mixed outcome gamble with the worst possible win. The results clearly supported the evoked reference set account.

Not only does the evoked reference set account naturally explain the full set of attractiveness ratings, it also casts doubt on the claim that the effect is counter-normative. Interpreting the subjective, bounded attractiveness scale requires implicitly, if not explicitly, answering the question of what the range of outcomes might be. With only a single gamble to evaluate, participants can only answer the question in rough and uncertain terms. They appear to make a parsimonious inference: If the lone gamble involves only wins, participants assume a reference set of gambles involving only wins. And if the gamble involves wins and losses, they assume a reference set of gambles involving wins and losses. In this respect, the observation that a "lose $0" outcome leads to higher ratings than a "win $0" outcome is especially telling. We showed in Experiment 3 that these logically equivalent descriptions provide relevant information to judges: Speakers are more likely to describe a $0 outcome as "win $0" (rather than "lose $0") when the other gambles in a reference set involve primarily wins as opposed to when the other gambles in the set involve a mixture of wins and losses.

However, in the model depicted in Figure 1, there are two ways in which evoked reference sets can influence subjective ratings. They might not just influence scale interpretation, but also subjective value. While the scale interpretation route is both necessary and normative, changes in subjective value are psychologically plausible but not compatible with standard normative models of choice. We attempted to determine the extent to which each route is affecting ratings by studying choices, which eliminate the scale interpretation route. If choices are influenced by adding the small loss to the standard gamble, then this effect cannot be attributed to scale interpretation and is likely due to effects of evoked reference sets on

subjective value. This, in turn, would suggest that inflated subjective value is influencing ratings, as well. Two of four choice experiments reported by Bateman et al. (2007) and Slovic et al. (2002) yielded evidence for a small loss effect.

In the present studies, while the effects of gamble on attractiveness ratings were robust and consistent, the effects on choices were relatively weak and often statistically non-significant. Indeed, we were only able to replicate the small-loss effect on choices in one of our four experiments examining choice (Experiment 5). Nonetheless, even when effects on choices were not significant, they were always in the direction predicted by evoked reference sets. We also found that participants consistently preferred the gamble with the "win $0" outcome description to a sure amount less often than they preferred the "lose $0" gamble to a sure amount. Thus, our data suggest that evoked reference sets have some influence on choice, but the effect may be relatively small and inconsistent. These results are interesting not just with respect to choices per se, but also their implications for attractiveness ratings. Because the choice task eliminates scale interpretation, the effects on choice indicate that evoked reference sets influence subjective value. However, in the ratings task, both scale interpretation and inflated subjective value are potential routes to a small loss effect, and the effects on ratings are large and consistent, suggesting that much of the work is being done by scale interpretation.

The current evoked reference set account is consistent with Kahneman and Miller's (1986) norm theory. They speculated that the standard gamble and the small-loss gamble might evoke different norms (analogous to reference sets). The present paper complements and refines this suggestion. First, we report empirical evidence that distinguishes a reference set explanation of the small-loss effect from the competing affective evaluability account (Experiment 2). Second, we show that the reference sets evoked by different gamble descriptions are well-attuned

to the effects of reference sets on speakers' gamble descriptions (Experiment 3). Finally, we distinguish between psychologically and normatively distinct pathways – scale interpretation and subjective value – through which reference sets may affect evaluation, and we assess their contributions through a comparison of ratings and choice.

The idea that evoked reference sets are important to phenomena involving judgments about isolated objects or events appears to be widely applicable. Beyond norm theory, recall Leong et al.'s (2016) finding that, in a between-subjects design, the number 2 was rated as smaller than the number -2. This is because the number 2 evokes a reference set of natural numbers (relative to which 2 is small), whereas -2 evoked a reference set of all integers (relative to which -2 falls in the middle). McKenzie and Soll (1996) discussed the possible role of evoked reference sets to account for which base rates are (not) used when making judgments in Bayesian inference tasks, and Gigerenzer, Hoffrage, and Kleinbölting (1991) argued that participants' under- and overconfidence in their beliefs as to which of two cities was larger depended on the different reference sets that were evoked by asking different types of questions.

Figure 1 distinguishes between a gamble's objective components (probabilities and outcomes) and how it is described (e.g., "win 0" vs. "lose 0"). Both components convey information to the judge about a reasonable reference set when providing a subjective rating, such as attractiveness. Information conveyed by the objective components is consistent with the "options as information" model (Sher & McKenzie, 2014), which provides a rational analysis of some apparently counter-normative context effects in human decision making. Researchers often compare decisions and evaluations across different choice sets, and regard inconsistent ordering across contexts as evidence for irrationality. However, when – as is often the case in these studies – the natural space of options is poorly known, participants may reasonably treat

the choice set as a sample from this space. Different choice sets may then lead to different inferences, which may in turn lead to different preferences. For example, joint-separate reversals occur when option A is rated higher than B when each is evaluated separately (i.e., between-subjects), but B is rated higher than A when evaluated jointly (i.e., within-subjects; e.g. Hsee, 1996). However, Sher and McKenzie (2014) showed that these findings are accounted for by a rational model in which judges learn from the presented options (about the distribution of attribute values) and update their preferences accordingly: Participants drew markedly different inferences from different (separate and joint) evaluation sets, and, when these inferences were presented as background information to a different group of participants, they sufficed to reproduce the joint-separate effect. A similar analysis has been applied to apparently intransitive choice behavior, in which choices made in multiple pairwise contexts are not compatible with a single underlying preference order (Müller-Trede, Sher, & McKenzie, 2015), and the phenomenon of asymmetric dominance, in which the addition of a third inferior "decoy" option to a "core" two-option choice set systematically alters preferences over the core options (Sher, Müller-Trede, & McKenzie, in preparation; see also Kamenica, 2008; Prelec, Wernerfelt, & Zettlemeyer, 1997; Ratneshwar, Shocker, & Steward, 1987). In these tasks, as in the gamble evaluation task, participants draw different -- and reasonable -- inferences depending on the specific options presented to them, and their expressed preferences and attitudes reflect this.

The influence of the gamble's description in Figure 1, on the other hand, is an example of "information leakage", a normative framework we have developed to explain some framing effects (Sher & McKenzie, 2006, 2008, 2011; see also Keren, 2007; McKenzie, 2004; McKenzie, Liersch, & Finkelstein, 2006; McKenzie & Nelson, 2003; Teigen & Karevold, 2005). The approach uses conversational pragmatics to shed light on why logically equivalent

utterances result in different listener behavior (Grice, 1975; Hilton, 1995; Schwarz, 1994). Generally, the fact that a speaker (e.g., experimenter) chooses to describe an option or outcome in a particular way (e.g., "win 0" vs. "lose 0", or beef that is "85% lean" vs. "15% fat") can convey relevant information to listeners. For example, a new medical treatment that leads to more survivors than other treatments is relatively likely to be described as leading to a "50% survival rate" (rather than a "50% mortality rate"). That is, the distribution of other treatments' efficacy influences a speaker's choice of description. Furthermore, listeners are sensitive to the speaker's choice of description. When a new treatment is described as having a "50% survival rate", listeners are relatively likely to infer that other treatments lead to lower survival rates. That is, describing a treatment in terms of its "survival rate" triggers an inference to a reference set of less effective treatments.

When seemingly inconsequential changes to the decision environment affect judgments and decisions, the result is usually construed as evidence of irrationality. However, it is necessary to rule out the possibility that the change in context provides information that a rational actor would utilize (Sher & McKenzie, 2008, 2011), and it is well known that decision makers can be sensitive to even subtle changes in context (e.g., Payne, Bettman, & Johnson, 1993). In this spirit, the present model distinguishes between normative and non-normative ways in which adding a small loss to a gamble can increase its attractiveness when evaluated in isolation. Data from our five experiments suggest that the higher attractiveness ratings for the dominated, small-loss gamble make sense, in part, because participants are making parsimonious inferences about the range of outcomes given the single gamble presented to them. This, in turn, influences the interpretation of the subjective scale in reasonable ways. Data from choices, however, indicate that the evoked context can also influence subjective value, though this non-

normative effect appears to be less robust in the present task. Adding a possible small loss does not just make a gamble slightly worse. It also conveys information about the task-relevant comparison set, relative to which the same gamble may emerge as far better.

References

Bateman, I., Dent, S., Peters, E., Slovic, P., & Starmer, C. (2007). The affect heuristic and the attractiveness of simple gambles. Journal of Behavioral Decision Making, 20, 365-380.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. Psychological Methods, 4, 243-249.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506-528.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (eds.), Syntax and Semantics Volume 3: Speech Acts. New York: Academic Press.

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. Psychological Bulletin, 118, 248-271.

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

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. Psychological Review, 93, 136-153.

Kamenica, E. (2008). Contextual inference in markets: On the informational content of product lines. American Economic Review, 98, 2127-2149.

Keren, G. (2007). Framing, intentions, and trust–choice incompatibility. Organizational Behavior and Human Decision Processes, 103, 238-255.

Leong, L. M., McKenzie, C. R. M., Müller-Trede, J., & Sher, S. (2016). When and why is 9 > 221? Reference sets evoked by the stimulus, rating scale, and elicitation method. Manuscript in preparation.

McGraw, A. P., Larsen, J. T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. Psychological Science, 21, 1438-1445.

McKenzie, C. R. M. (2004). Framing effects in inference tasks – and why they are normatively defensible. Memory & Cognition, 32, 874-885.

McKenzie, C. R. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. Psychological Science, 17, 414-420.

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 and Review, 10, 596-602.

McKenzie, C. R. M., & Soll, J. B. (1996). Which reference class is evoked? Behavioral and Brain Sciences, 19, 34-35.

Müller-Trede, J., Sher, S., & McKenzie, C. R. M. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. Decision, 2, 280-305.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The adaptive decision maker. Cambridge: Cambridge University Press.

Prelec, D.,Wernerfelt, B., & Zettlemeyer, F. (1997). The role of inference in context effects: Inferring what you want from what is available. Journal of Consumer Research, 24, 118-125.

Ratneshwar, S., Shocker, A. D., & Steward, D. W. (1987). Toward understanding the attraction effect: The implications of product stimulus meaningfulness and familiarity. Journal of Consumer Research, 13, 520-533.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. Advances in Experimental Social Psychology, 26, 123 – 162.

Sher, S. & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. Cognition, 101, 467-94.

Sher, S., & McKenzie, C. R. M. (2008). Framing effects and rationality. In M. Oaksford & N. Chater (Eds.), The probabilistic mind: Prospects for Bayesian cognitive science (pp. 79-96). Oxford: Oxford University Press.

Sher, S., & McKenzie, C. R. M. (2011). Levels of information: A framing hierarchy. In G. Keren (Ed.), Perspectives on framing (pp. 35-64). Psychology Press - Taylor & Francis Group.

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

Sher, S., Müller-Trede, J., & McKenzie, C. R. M. (2016). Asymmetric information in asymmetric dominance: The informational value of worthless options. Manuscript in preparation.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 397-420). Cambridge: Cambridge University Press.

Teigen, K. H., & Karevold, K. I. (2005). Looking back versus looking ahead: Framing of time and work at different stages of a project. Journal of Behavioral Decision Making, 18, 229-246.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. Cognitive Science, 38, 599-637.

## Author Note

This research was supported by National Science Foundation Grants SES-0820553, SES-1060270, and a Scholar Award (to the second author) from the James S. McDonnell Foundation.
Correspondence should be addressed to Craig R. M. McKenzie, UC San Diego, 9500 Gilman Drive – MC 0553, La Jolla CA 92093-0553. E-mail: cmckenzie@ucsd.edu.

## Footnote

1. In Birnbaum's (1999) original study, participants in a between-subjects design rated the number 9 as larger than the number 221. Consistent with the present analysis, Birnbaum speculated that 9 evokes a reference set of one-digit numbers whereas 221 evoked a reference set of three-digit numbers. However, Leong et al. (2016) showed that the "9 > 221 effect" is more complicated: It critically depends on the numerical values included in the rating scale itself, which, in this task, may also influence the evoked reference set to which the target is compared.

Appendix: Experiment 3 Stimuli

After a short page of general instructions, participants read the following on page 2:

> In this experiment, you will read about a number of possible gambles. Then one of the gambles will be selected. You will be asked to describe this selected gamble to someone else.
>
> The Table on the next page lists all of the possible gambles. For each gamble, there are two potential *outcomes*. Each outcome is some amount of money the person playing the gamble could win or lose. If the amount of money is positive, this means the person wins that amount of money. If the amount is negative, this means the person loses that amount of money.
>
> In every gamble, each of the two monetary outcomes has a certain *probability*. This probability indicates how likely it is that a person playing this gamble will have this monetary outcome.
>
> For example, please look at **Gamble 4** in the Table on the next page. In this gamble, a person will win $18 with probability 1/4, and will lose $5 with probability 3/4.
>
> In other words, 1 out of every 4 times a person plays Gamble 4, they will win $18, and the rest of the time they play Gamble 4 they will lose $5.
>
> *Now please turn the page, and carefully read through all of the12 gambles listed in the Table.*
>
> *If you don't understand anything about the Table, feel free to turn back to this page to re-read the explanation above.*

On the third page was the table listing 12 gambles and further instructions. Below is the table presented to the "Win/Win" participants. The table presented to the "Win/Lose" participants was identical except that all of the non-zero values in the Outcome 2 column were negative (i.e., were losses).

Table of Gambles

| **Gamble** | **Probability of Outcome 1** | **Outcome 1** | **Probability of Outcome 2** | **Outcome 2** |
|---|---|---|---|---|
| 1 | 2/5 | $10 | 3/5 | $2 |
| 2 | 7/14 | $5 | 7/14 | $12 |
| 3 | 2/3 | $20 | 1/3 | $3 |

| | | | | |
|---|---|---|---|---|
| 4 | 1/4 | $18 | 3/4 | –$5 |
| 5 | 19/40 | $8 | 21/40 | $11 |
| 6 | 90/100 | $0 | 10/100 | $17 |
| 7 | 14/16 | $3 | 2/16 | –$8 |
| 8 | 3/8 | $4 | 5/8 | $1 |
| 9 | 1/12 | $3 | 11/12 | $4 |
| 10 | 7/36 | $9 | 29/36 | $0 |
| 11 | 5/9 | $4 | 4/9 | $3 |
| 12 | 3/20 | $16 | 17/20 | $2 |

*When you have finished reading all the gambles in the Table, please turn to the next page.*

On the next page, participants were presented with the following (though for half the participants the options to be circled had "lose" on top and "win" on the bottom):

> Now imagine you had to describe **Gamble 10** to a friend, who will then have to decide whether to play this gamble.
>
> Please review Gamble 10 in the Table on the previous page, and then complete the following description as appropriate.

"In Gamble 10, there is a _____/_____ chance that you will win / lose $_____ ,
*(write #'s)* *(write #)*

*(circle one)*

and there is a _____/_____ chance that you will win / lose $_____ ."
*(write #'s)* *(write #)*

*(circle one)*

Figure Captions

Figure 1: Proposed model of how a single gamble is evaluated. Both the content and the framing of the gamble affect its subjective value and also evoke a reference set. When the gamble is evaluated on an underspecified subjective scale, the reference set must be used to interpret the scale (top solid arrow), and an inferior evoked reference set leads to higher ratings. The model includes a second optional path (broken arrow) whereby reference sets may directly induce contrast effects on subjective value, potentially leading to scale-independent effects.

Figure 2: Experiment 1: Attractiveness ratings as a function of the different gambles. Standard error bars are shown.

Figure 3: Experiment 1: Percentage of participants choosing to play each gamble rather than receive $2 for sure. Standard error bars are shown.

Figure 4: Experiment 2: Attractiveness ratings as a function of the different gambles. Standard error bars are shown.

Figure 5: Experiment 2: Percentage of participants who preferred the gamble to either a sure gain of $2 (left two bars) or a sure loss of $2 (right two bars). Standard error bars are shown.

Figure 6: Experiment 4: Percentage of participants who preferred the gamble to receiving $2 for sure as a function of the gamble. Standard error bars are shown.

Figure 7: Experiment 5: Attractiveness ratings (top panel) and percentage of participants who preferred the gamble to $3 for sure (bottom panel) as a function of the gamble and whether the attractiveness question or choice was first. Standard error bars are shown.

Figure 1

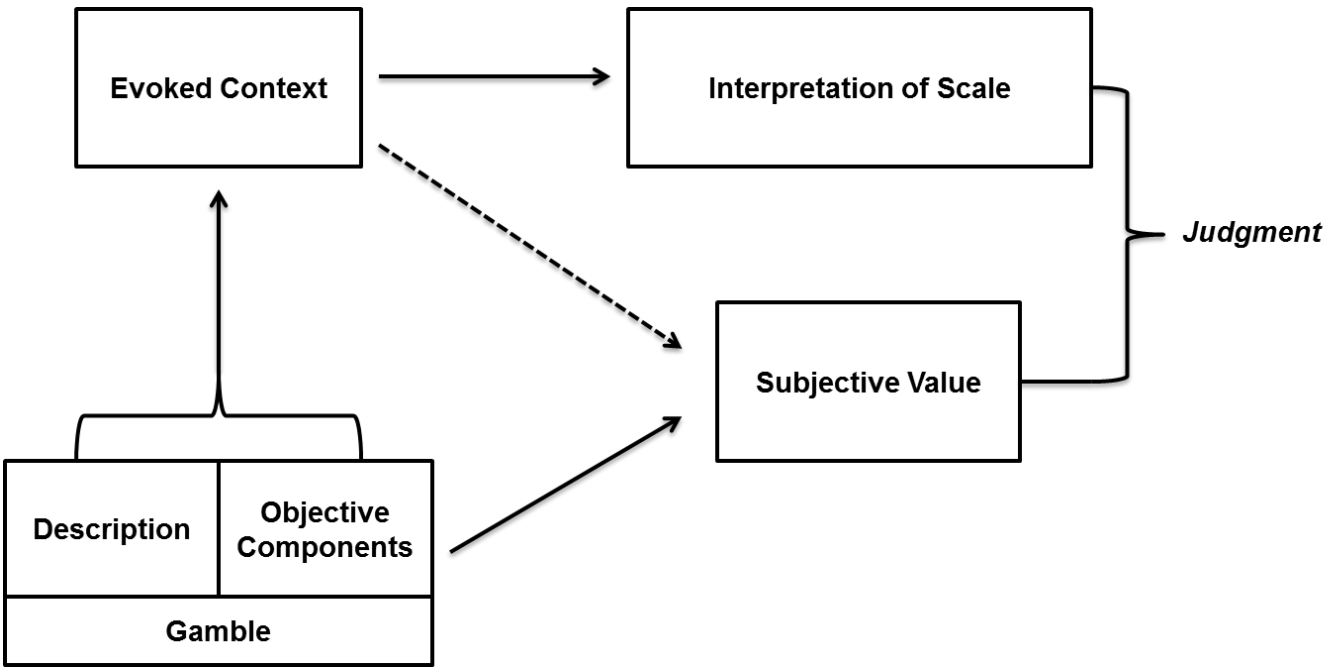

Evoked Context
Interpretation of Scale
Judgment
Subjective Value
Description
Objective Components
Gamble

Figure 2

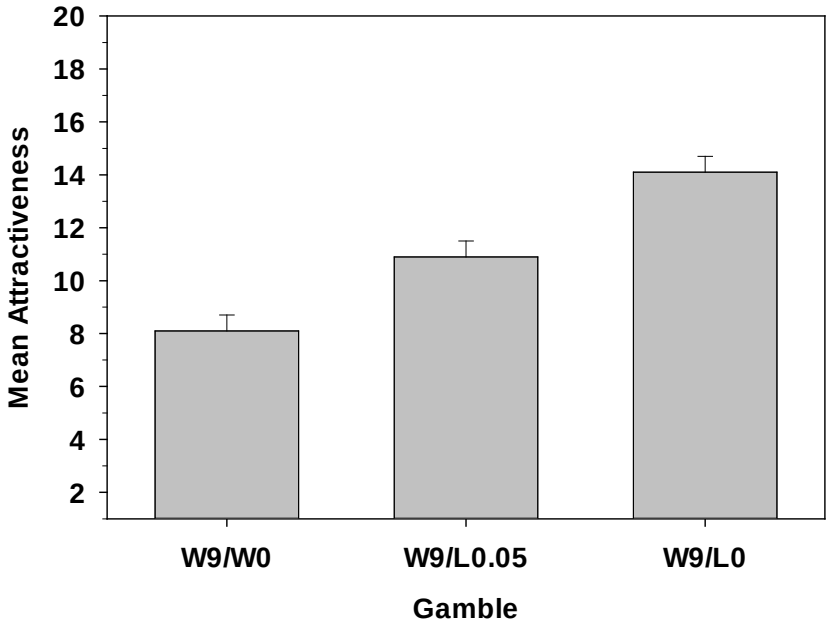

20
18
16
14
12
10
8
6
4
2
Mean Attractiveness
W9/W0
W9/L0.05
W9/L0
Gamble

Figure 3

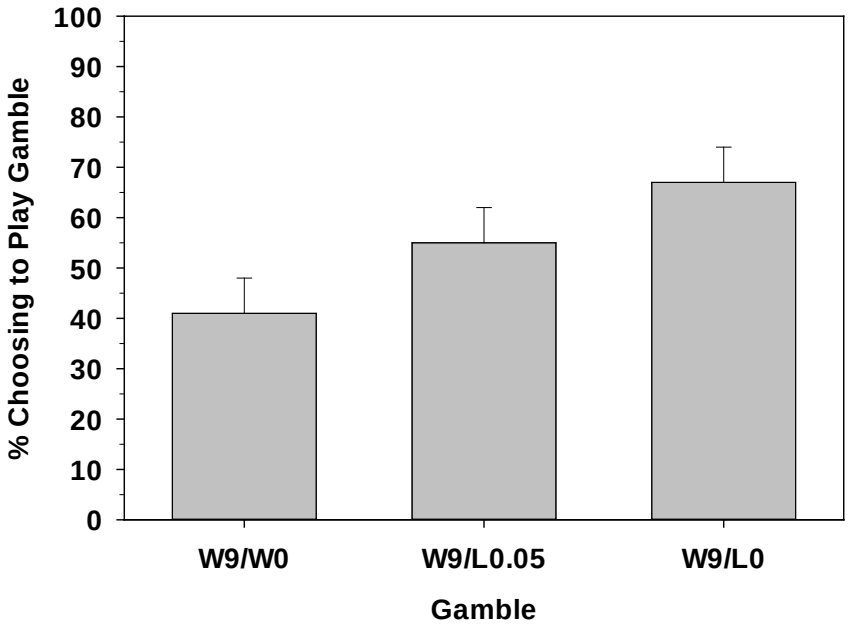

100
90
80
70
60
50
40
30
20
10
0
% Choosing to Play Gamble
W9/W0
W9/L0.05
W9/L0
Gamble

Figure 4

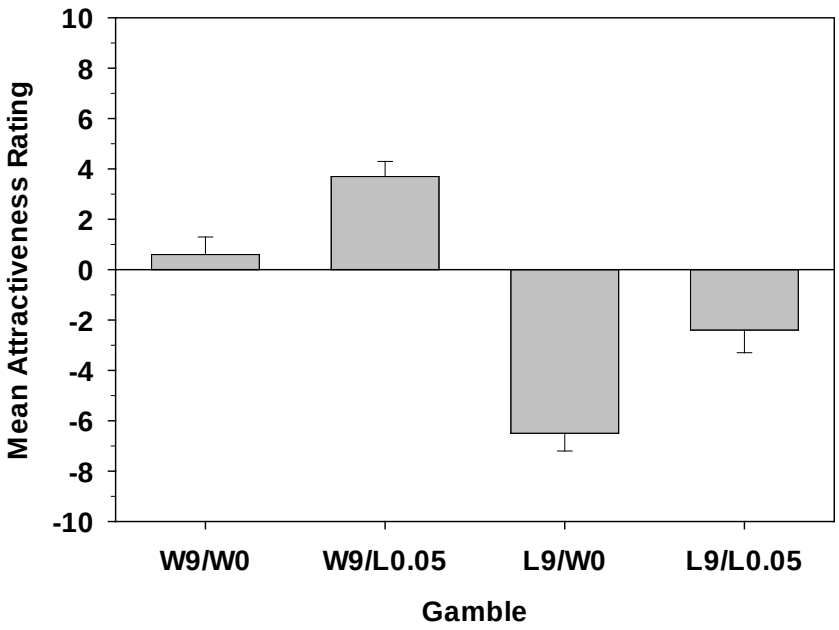

10
8
6
4
2
0
-2
-4
-6
-8
-10
Mean Attractiveness Rating
W9/W0
W9/L0.05
L9/W0
L9/L0.05
Gamble

Figure 5

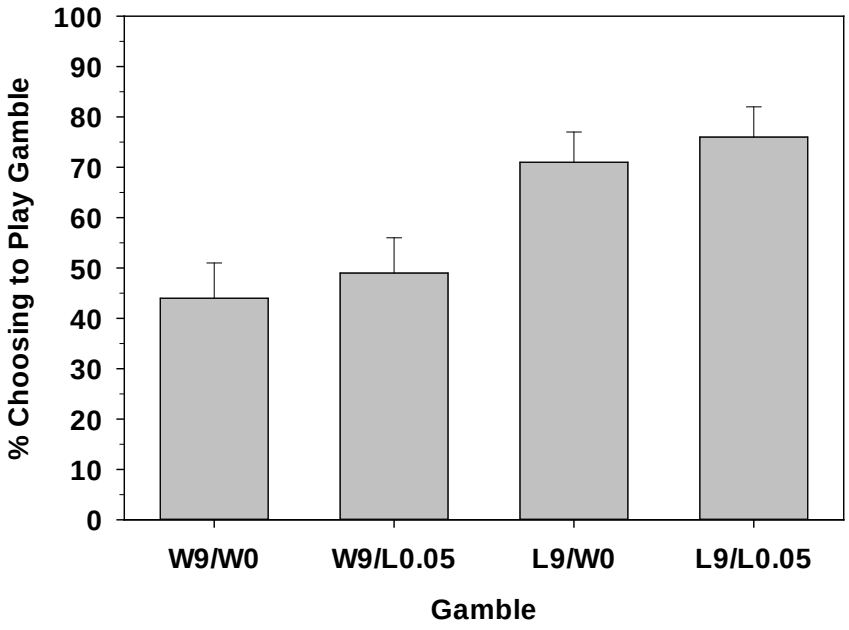

100
90
80
70
60
50
40
30
20
10
0
% Choosing to Play Gamble
W9/W0
W9/L0.05
L9/W0
L9/L0.05
Gamble

Figure 6

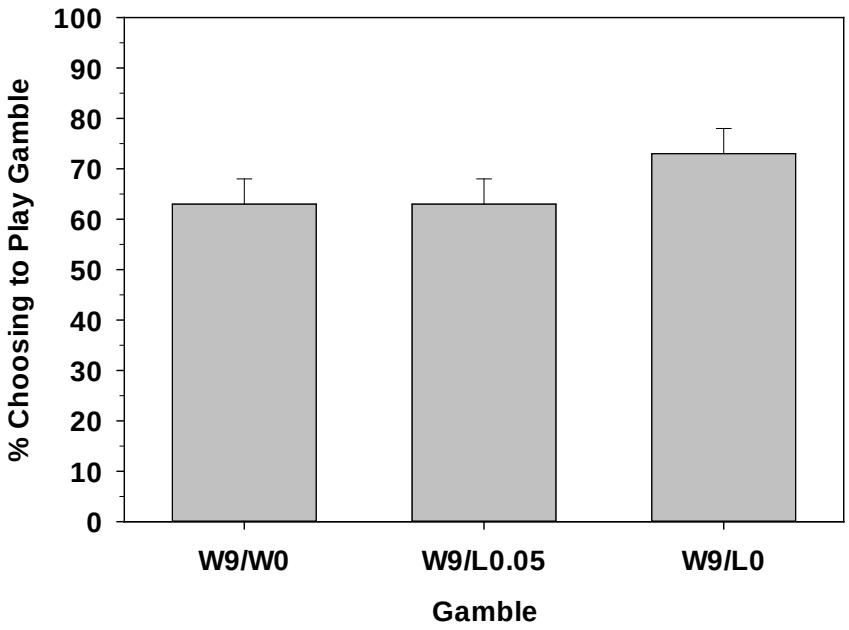
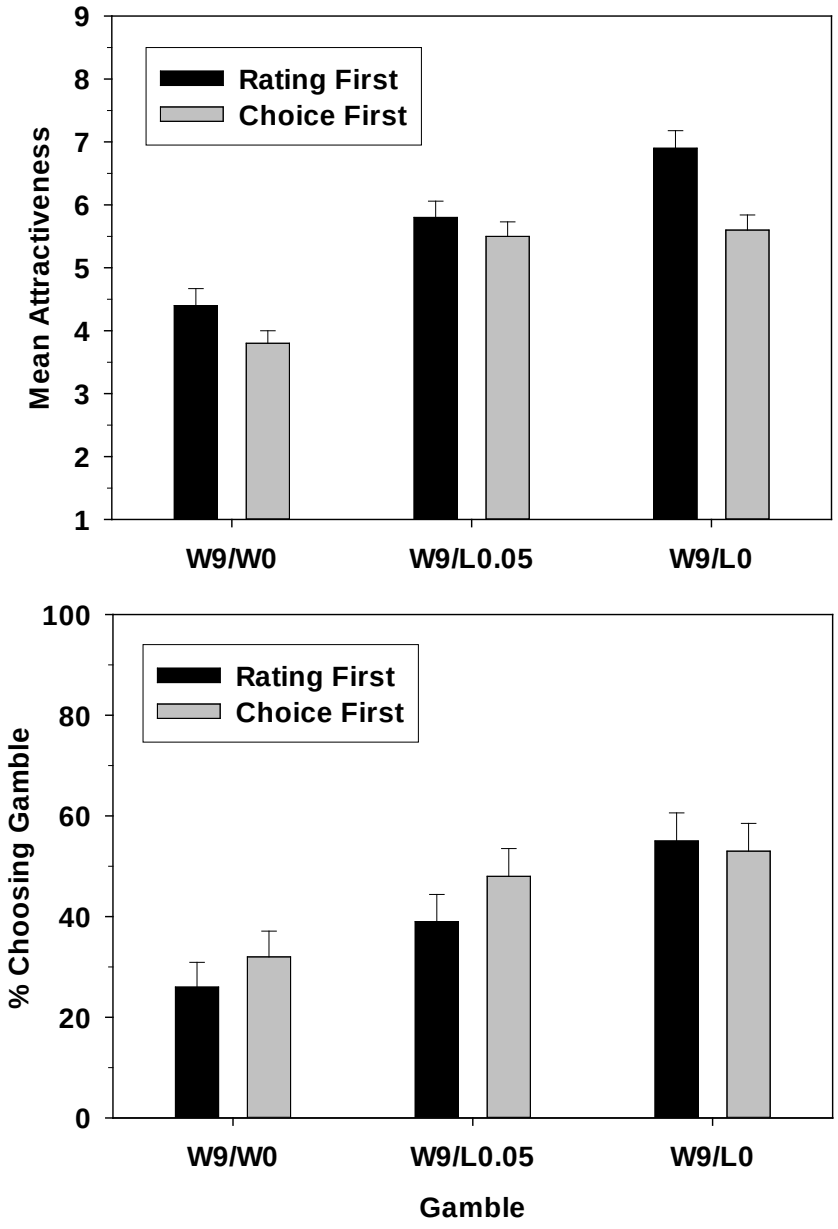

100
90
80
70
60
50
40
30
20
10
0
% Choosing to Play Gamble
W9/W0
W9/L0.05
W9/L0
Gamble

Figure 7


Mean Attractiveness
9
8
7
6
5
4
3
2
1
Rating First
Choice First
W9/W0
W9/L0.05
W9/L0
% Choosing Gamble
100
80
60
40
20
0
Rating First
Choice First
W9/W0
W9/L0.05
W9/L0
Gamble