

**Evaluations Are Inherently Comparative, But Are Compared To What?**

Minah H. Jung<sup>1</sup>, Clayton R. Critcher<sup>2</sup>, Leif D. Nelson<sup>2</sup>

<sup>1</sup>New York University, <sup>2</sup>University of California, Berkeley

Authors' Note

Minah H. Jung and Clayton R. Critcher equally contributed to this research as joint lead authors.

Minah H. Jung, Leonard N. Stern School of Business, New York University; Clayton R.

Critcher, Haas School of Business, University of California, Berkeley; Leif D. Nelson, Haas

School of Business, University of California, Berkeley. We thank Chengyao Sun, Ingrid Adams,

Phoebe Wong, and Xinyu Wei for their support for this research. All raw data and related coding

information underlying all findings have been shared on OSF

([https://osf.io/cdxsw/?view\\_only=ff4bd53e394c4b8c978b9f785ee14e7d](https://osf.io/cdxsw/?view_only=ff4bd53e394c4b8c978b9f785ee14e7d)).

Correspondence concerning this article should be addressed to Minah H. Jung, Leonard N. Stern

School of Business, New York University, 40 West 4<sup>th</sup> Street, Tisch Hall 912, New York, NY

10012. E-mail: [minah.jung@stern.nyu.edu](mailto:minah.jung@stern.nyu.edu).

### Abstract

Understanding how objective quantities are translated into subjective evaluations has long been of interest to social scientists, medical professionals, and policymakers with an interest in how people process and act on quantitative information. The theory of decision by sampling proposes a comparative procedure: Values seem larger or smaller based on how they rank in a comparison set, the *decision sample*. But what values are included in this decision sample? We identify and test four mechanistic accounts, each suggesting that how previously encountered attribute values are processed determines whether they linger in the sample to guide the subjective interpretation, and thus the influence, of newly encountered values. Testing our ideas through studies of loss aversion, delay discounting, and vaccine hesitancy, we find strongest support for one account: Quantities need to be subjectively evaluated—rather than merely encountered—for them to enter the decision sample, alter the subjective interpretation of other values, and then guide decision making. Discussion focuses on how the present findings inform understanding of the nature of the decision sample and identify new research directions for the longstanding question of how comparison standards influence decision-making.

*Keywords:* decision by sampling, risky decision making, delay discounting, vaccine hesitancy, comparison standards

### **Evaluations Are Inherently Comparative, But Are Compared To What?**

In order to make good decisions, people need to evaluate the available options. In some cases, choices become easy-to-define (even if difficult-to-calculate) exercises in maximization. Such problems require complex math, not complex representations of value. For example, when bees determine what route to take when searching for food, they are attempting to maximize the calories found minus the calories expended. And bees, like computers, seem optimized to solve objectively solvable optimization problems of this sort (Real, 1991, 1996).

Other choices—particularly those that involve tradeoffs—are more complex. In deciding whether to upgrade a smartphone or an airline seat, one must consider whether such enhancements are “worth it.” This requires people to translate objective changes in attribute values into subjective assessments of value. Understanding how the mind arrives at such subjective valuation has been a core project for psychologists, economists, and others interested in judgment and decision making (Bernoulli, 1954; Edwards, 1954; Kable & Glimcher, 2007; Kahneman & Tversky, 1979; Parducci, 1968; Stewart et al., 2006).

Prospect theory emphasized—among other innovations—that valuation is inherently comparative (Kahneman & Tversky, 1979). Two airlines currently offering equivalent legroom may find themselves with differently satisfied customers if one airline arrived at that outcome by adding an inch whereas the other subtracted an inch from a previous seat layout. That is, identical current values may be evaluated differently when they are compared to different reference points. Numerous framing problems demonstrate how even arbitrary reference points can produce sizable shifts in valuation (Kühberger, 1998; Levin et al., 1998; Tversky & Kahneman, 1985).

Psychologists have long appreciated that human judgment is inherently comparative. For example, a fourteen-year-old student is young in a college classroom and old in an elementary school. More generally, comparisons feature prominently in many decision-making models (Busemeyer & Townsend, 1993; Kahneman & Miller, 1986; Tversky & Koehler, 1994). This reality leads judgments and decisions to seem unstable, even if theoretical accounts can explain why such variation is not arbitrary (Allik & Tuulmets, 1991; Ariely et al., 2003; Garner, 1962; Laming, 1984, 1997; Miller, 1956; Shiffrin & Nosofsky, 1994).

More recently, psychologists have argued that even evaluations that seem to be made in isolation may themselves be comparative in nature. Namely, one particular theory, decision by sampling (DbS), has been proposed to suggest how comparison processes are even more deeply entrenched in valuation (Stewart et al., 2006; see, Noguchi & Stewart, 2018, for a crucial extension of DbS) than previous theories anticipated. The theory stands in contrast to a “value-first” approach to decision making (Vlaev et al., 2011), one in which there exists a preexisting, direct mapping in the mind from a specific attribute value to a subjective valuation. With decision by sampling, attribute values feed into *comparison-based decision making* (Vlaev et al., 2011), such that a stimulus value informs judgments and decisions based on how that attribute value ranks in a set of relevant comparison values. In this way, subjective valuations are fundamentally and necessarily comparative.

For example, an airline passenger who looks down at her 15 inches of legroom may evaluate the airline’s generosity by calling to mind a sample of previously encountered values against which this new value is compared. If on recent flights she experienced legroom of 12, 13, 13, 13, and 16 inches, her current seating may seem spacious—better than 4/5 of those in the decision sample. But were she to be returning to the skies after a twenty-year hiatus, her decision

sample may be 14, 18, 18, 19, and 19 inches. For her, the legroom may be experienced negatively as a 1/5—better than only one of the five values in her decision sample. The theory also anticipates that the same objective shift in an attribute will be experienced differently by different people to the extent it has different effects on the attribute's rankings in their (potentially idiosyncratic) decision samples. Adding an extra two inches of legroom should have more of an impact on the former traveler (for whom the legroom's ranking changes) than for the latter one (for whom the new legroom would have the same low ranking).

Decision by sampling complements a rich tradition rooted in psychophysics that has examined how the evaluation of stimuli depends on the presence of other stimuli in the judgment context (e.g., Sherif et al., 1958). In some cases, other stimuli help to define a category or a context that exerts an assimilative pull on how target stimuli are interpreted or represented in memory (Crawford et al., 2000; Huttenlocher et al., 2000; Neisser, 1976; Liberman et al., 1967), especially for those from collectivistic cultures who are prone to process stimuli in light of other stimuli presented in that context (Duffy & Kitayama, 2007; Oyserman & Lee, 2008). But in other cases, contextual stimuli become standards of comparison that help to frame target stimuli as large or small, substantial or insubstantial (see Bless & Schwarz, 2010, for a review of these dual roles).

For example, range theory (Volkman, 1951) proposed that perceivers make sense of target stimuli based on where they fall in the range of stimuli that define a context (i.e., how far the target stimuli are from the most extreme exemplars). Consistent with this, Mellers and Cooke (1994) found that apartment hunters who considered options that varied a lot in their rent (vs. were all similarly priced) were less enticed by a \$50 discount in rent to accept a drawback (e.g., a longer commute). Given the range of rents those participants saw, \$50 just didn't seem to be all

that much. Range-frequency theory (Parducci, 1963, 1983) went a step further in positing that the density (i.e., frequency of occurrence) of exemplars along that range also shapes understanding. For example, two cities seem further apart when there are more other cities in between them (Thorndyke, 1981). Decision by sampling also suggests judgments are made relative to a distribution of comparative stimuli, but it is only a target's ranking in the decision sample—not its position in the range of such values—that the theory identifies as crucial.

Empirical support for DbS has been substantial and uncovered in various domains. Numerous investigations have documented how people's judgments and decisions are seemingly informed by comparisons with values to which they are recently or chronically exposed (Olivola & Sagara, 2009; Stewart et al., 2006; Walasek & Stewart, 2015). Although there is a substantial understanding of *how* decision samples guide evaluation and decision making, less is known about *which* quantitative values enter such samples. Without a better answer to the latter question, the promise of DbS will be unfortunately limited.

Understanding which values become standards for comparison is important, both in a local sense (by addressing a hole in a relatively young psychological theory), and a global sense (given the scope of human experience to which theories of decision making apply). As it stands, social scientists, practitioners, and policymakers tend to look to prospect theory for insight into how attribute values are subjectively interpreted. As a result, prospect theory's influence is easily seen: in sociology (Steinacker, 2006), law (Duffy, 2004; Guthrie, 2003; Korobkin, 2012), economics (Barberis, 2013), as well as other disciplines. Prospect theory underpins interventions that influence jury decision making in civil trials (Jolls & Sunstein, 2006; McCaffery et al., 1995), the prevention of HIV (Mcdermott, 1998), and the encouragement of retirement savings (Thaler & Benartzi, 2004). Social psychologists in particular have turned to prospect theory to

understand cooperation and competition in social dilemmas (De Dreu & McCusker, 1997), to make sense of white and Black Americans' diverging perceptions of racial progress (Eibach & Keegan, 2006), and to design interventions that encourage preventative healthy behaviors (Detweiler-Bedell & Detweiler-Bedell, 2016; Rothman & Salovey, 1997), among other varied applications.

As reviewed above, decision by sampling has the potential to extend even further than prospect theory. By identifying comparison processes as playing a more central role, DbS provides an even more comprehensive model of decision making. Subjective valuation is arrived at not merely by how an attribute compares against a single reference point—something that has been hard to identify in certain decision making domains (e.g., political science; Mercer, 2005)—but against all values in a decision sample. As a result, DbS can anticipate and explain certain phenomena that prospect theory either cannot or has to merely posit (Stewart et al., 2006). And indeed, the immense promise of DbS is already being seen in its application in different domains.

More generally, rank-based comparison models have been used to explain the imperfect relationship between income and happiness (Boyce et al., 2010), how feedback influences employee effort (Gill et al., 2019), and who is more interested in status-signaling goods (Walasek & Brown, 2015). Even outside of more typical psychological foci, DbS has helped to explain drinkers' perceptions of the riskiness of their own alcohol consumption (Wood et al., 2012) as well as people's impressions of their own depression symptom severity (Melrose et al., 2012). Vlaev et al.(2011) argued that comparison-based decision theories like decision by sampling have the potential to be broadly influential by requiring the updating of philosophical theories, classical economic assumptions, and even the manner in which medical decision

making should unfold. That said, the comprehensiveness of DbS as a theory of how subjective valuation originates (i.e., through comparisons) is both a reflection of its broad potential but also an explanation for why that potential has yet to be fully realized. That is, the theory does not specify *which* values enter decision samples to influence the interpretation of newly encountered quantitative attributes. As a result, a substantive advance on this front would be important not merely to DbS and to comparison-based decision theories more generally, but to the broad range of human activity to which these theories can speak.

Stewart et al. (2006) say recent exposure likely matters, but “similarity and background knowledge will surely play a role as well” (p. 4). Sometimes, this refers to *categorical* similarity: Evaluation of salaries and car prices may rely on comparisons with other salaries and car prices, respectively, but not with each other (Rablen, 2008; see also Hounkpatin et al., 2016). It can also mean *value* similarity: A person considering a \$649 mobile phone is more likely to use a \$689 phone as a relevant comparison than a \$999 phone (Brown & Stewart, 2005; Qian & Brown, 2005). Other research suggests value *distinctiveness* plays a role: If one is exposed to many highly similar values, the chance that any individual one of those values enters the decision sample may be reduced (Brown & Matthews, 2011; Brown et al., 2007; Tripp & Brown, 2016).

What unifies these previous efforts is their focus on trying to identify properties of attribute *values* that predict their inclusion in a decision sample. This approach is limited because, outside of experimental contexts, people see so many quantities that it becomes difficult to predict the resulting sample. We instead ask how the *processing* of values—whether passively through mere exposure or actively in the form of a cognitive procedure—influences their inclusion in a decision sample. Consider a recent paradigm used by Walasek and Stewart, (2015), who argued that decision by sampling can help to explain loss aversion. Loss aversion



describes the tendency for judgments, decisions, and experiences to be affected more by losses than equal-magnitude gains (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Social psychologists, like behavioral scientists more generally, appeal to loss aversion to explain various phenomena: why people make different decisions on behalf of their family members than their friends (Guassi Moreira et al., 2021), why people value goods more when they already possess them compared to when they are considering acquiring them (Kahneman et al., 1991), why professional basketball teams are counterproductively conservative in the final seconds of a close game (Walker et al., 2018), and even why those with optimistically mistaken expectations may experience lower well-being (de Meza & Dawson, 2021).

In Walasek and Stewart's (2015) research, participants were shown a series of lotteries and indicated which ones that they would play. The authors varied the distribution of losses and/or gains that defined such lotteries in an effort to change the distribution of the decision sample. Participants' loss aversion—their choices' greater sensitivity to changes in losses than changes in gains—changed just as decision by sampling would predict. But why?<sup>1</sup> We decompose different aspects of this manipulation to identify which one or more contribute to attribute values' inclusion in the decision sample. Participants were *exposed* to the lotteries, which included values to which they had to be *responsive*, requiring them to *evaluate* their attractiveness, and then ultimately make a *decision* about whether to take the risk. Precisely which one or more of these four steps accounts for the values' placement in the decision sample, thereby influencing decisions, is unclear. Answering this question—which can be posed not

---

<sup>1</sup> This paradigm has been the subject of recent controversy (André & de Langhe, 2021). When relying on variations on this paradigm, our own analyses will validate André and de Langhe's (2021) central concern and use a pair of analytic approaches that, in combination, are not subject to the same critiques (Walasek et al., 2021; Walasek & Stewart, 2021).

merely for loss aversion but other phenomena as well—should enable a more precise understanding of which values guide comparative decision making.

### **The Importance of How Values Are Processed**

To our knowledge, our studies reflect the first effort to examine how the way values are processed—as opposed to properties of the values themselves—leads them to be used as comparisons that guide decision-making. We posit and test four distinct processing mechanisms—exposure, responding, evaluation, and decision-making—that may be responsible for placing attribute values in a decision sample. On initial consideration, the question of what leads attribute values to linger in the decision sample might seem tantamount to the question of what makes such values memorable. But memory probes suggest that the contents of the decision sample are not simply recovered by recall measures (Walasek & Stewart, 2019). And to foreshadow, we will conceptually replicate these null effects. We will wait until later—once we are armed with a more solid understanding of what processing mechanism places values in the decision sample—to propose a revised understanding of how and when attribute values serve as critical comparisons.

**Exposure.** The exposure account anticipates that appearance in a decision sample follows from people's passive exposure to values. Mere exposure to attribute values will increase accessibility, and increased accessibility may move those values into the decision sample. The accessibility of an attribute value is determined by the chronicity and recency of exposure (Higgins, 1996). Previous research has shown that unobtrusive, or even subliminal, exposure to values can influence comparative numeric processing and quantity judgments (Greenwald et al., 2003; Kunde et al., 2003; Mussweiler & Englich, 2005). A similar process may underlie the mental creation of a decision sample: Incidental exposure to multiple numeric values may create

a locally relevant set to be used for forming subjective evaluations. This is the most inclusive account, one that casts the widest net for what attribute values will enter the decision sample.

That said, certain limits on when stimuli serve as comparison standards make it unlikely that exposure is sufficient. As one example, participants judged a weight as lighter if it followed interaction with a heavier object, but not if participants' interaction with the heavy object was purely incidental (Brown, 1953). It thus seems that people need to be more than exposed to a stimulus, but to have engaged with it in a more specific way, for it to enter the decision sample.

**Response.** Mere exposure to values is particularly passive. Instead, people may need to more actively focus on values in order for them to enter the decision sample. That is, instead of merely perceiving a value, people may need to focus on the value long enough and with sufficient depth that they actively respond in light of the value. To make that a minimal response—one that merely guarantees that the value has been focused on and encoded—we will ultimately operationalize this by asking participants to merely repeat the value back. The response account anticipates that this will place values in the decision sample.

**Evaluation.** It may not be sufficient to actively think about numbers for them to enter the decision sample. Instead, attribute values may enter the decision sample only once they are themselves subjectively evaluated. This link has intuitive appeal: Subjective valuations are achieved through comparisons, and such evaluated values may then form the set of values that guide future subjective valuations. For example, Schwarz et al. (1990) found that Germans judged beverages like wine and coffee to be more *typically German* if the German participants first estimated how often Germans drank vodka (an atypical drink for Germans) than if they first estimated vodka's calorie count. In other words, first evaluating the target on the dimension of interest led it to be used as a standard of comparison. The *evaluation* account thus sets a much

narrower criterion for entry into the decision sample than the exposure or response accounts because it requires active evaluation rather than minimal consideration.

**(Context-specific) decision.** Perhaps attribute values enter a decision sample not merely when they are subjectively evaluated, but instead when they have been subjectively evaluated in the context of the decision one is confronting. For example, someone casually browsing cars may be exposed to many car prices that they subjectively evaluate (e.g., “Wow, that one is really expensive!”). But it may be only the values one encounters while actively shopping—i.e., those about which one makes an actual *decision* to buy or not buy a car (e.g., “...which is why I will definitely not be buying that car”)—that are recruited when making subsequent decisions about whether to buy another car. By this account, the decision sample recruited in the service of a decision is determined by the values subjectively evaluated as part of making *that* decision in the past.

## Overview of Studies

We conducted experiments to determine whether one or more of these four processing mechanisms explained what attribute values enter the decision sample. The decision sample is not directly measurable but can be inferred indirectly by whether attribute values change decisions as decision by sampling anticipates. The first two studies disentangle the accounts by using a risky decision paradigm modified from that developed by Walasek and Stewart (2015). In Study 3, we ported our theory to a novel domain (patience) to see if the conclusions reached in one paradigm can be generalized. Finally, Study 4 aimed to replicate these findings in yet another context, one that is particularly socially relevant (vaccine hesitancy), while also more directly documenting the proposed process by which processing certain values influences decision making that relates to other values.

Across studies, we took several steps to enhance statistical power. First, in all studies, we measured our key dependent variable of interest using many trials. Second, we aimed to achieve large sample sizes. In each study, we recruited a sample size that would permit an average of at least 250 participants per condition. When funds allowed (or we suspected *a priori* that a specific prediction would be more nuanced), we exceeded this rule so that our average sample size was over 357 participants per condition. The relevant institutional review boards of the authors' universities approved all studies reported in this research. For all studies, we preregistered our sample size, design, and analysis plan. These preregistrations, along with full data and materials, can be found on the Open Science Framework:

[https://osf.io/a9362/?view\\_only=ff4bd53e394c4b8c978b9f785ee14e7d](https://osf.io/a9362/?view_only=ff4bd53e394c4b8c978b9f785ee14e7d).

### Study 1

Studies 1 and 2 modify Walasek and Stewart's (2015) loss-aversion paradigm to test what psychological processes are responsible for placing values in decision samples. In their paradigm, participants confront mixed gambles, indicate which ones they would accept, and thereby reveal a loss aversion coefficient. Whereas Walasek and Stewart varied the distribution of attribute values that defined the lotteries that participants saw and decided to accept or reject, our key manipulations instead encouraged some participants to engage with some attribute values in different ways—crucially, in several ways that did not require participants to actually make the decision of whether to accept or reject the lottery. Testing how this manipulation influences the degree of displayed loss aversion allowed us to examine what processes are responsible for placing attribute values in the decision sample.

In Study 1, all participants were exposed to the same gain (from \$6 to \$32) and loss values (from -\$20 to -\$6). But we varied how participants were led to process *some* of those

values—in particular, the larger gain values (from \$22 to \$32). Whereas some participants were merely exposed to these larger gains (narrow + exposure condition), the remainder were randomly assigned to one of three other conditions. One group responded to the larger gain values by retyping them (narrow + response condition), others had to subjectively evaluate the attractiveness of the lotteries defined by such values (narrow + evaluation condition), and still others made decisions about whether to accept or reject the lotteries with these larger gain values (wide condition). That is, we varied whether participants made decisions about a set of lotteries defined by a *narrow* (\$6 to \$20) or *wide* (\$6 to \$32) range of gains, but those in the narrow condition varied in how they engaged with the larger gains (\$22 to \$32): *exposure*, *response*, or *evaluation*. First, we expected to find that loss aversion was greater in the wide than the narrow + exposure condition. Second, we were particularly interested in the positioning of the narrow + response and narrow + evaluation conditions given they permit tests of what process or processes are responsible for placing attribute values in the decision sample.

Note that in Study 1, all participants were at least exposed to the full range of gain values (from \$6 to \$32). This is because before conducting Study 1, we conducted Supplemental Study A, which tested and failed to find support for the exposure account, the idea that merely being exposed to an attribute value would be sufficient to place it in the decision sample. The study found that participants who confronted mixed lotteries defined by a wide (\$6 to \$32) instead of a narrow (\$6 to \$20) range of gains did indeed show greater loss aversion. Decision by sampling anticipates this finding because the same objective change in gain values (e.g., from \$6 to \$20) more changes the normalized (or percentile) rank in the set of narrow gains (for which \$6 and \$20 are the two extremes) than in the set of wide gains (in which a shift from \$6 to \$20 passes through only some of the attribute values). This makes those considering the narrow gain range

more sensitive to the same-sized changes in gains, which diminishes loss aversion (the relative sensitivity to changes in losses as opposed to gains). But for those who made decisions about lotteries defined over the narrow range of gains (i.e., \$6 to \$20), their display of loss aversion was not magnified by their being exposed to the full range of gain values seen by those in the wide condition (i.e., \$22 to \$32). Such exposure was accomplished by showing these *narrow + exposure* participants the possible gain value of each lottery not on a number line that merely extended from \$6 to \$20 (as in the narrow condition), but on a number line whose upper bound was \$32 (and on which the numbers \$22, \$24, \$26, \$28, \$30, and \$32 were explicitly labeled). Guided by this finding that mere exposure is insufficient to place values in the decision sample, Study 1 uses such a narrow + exposure condition as the baseline control.

## Method

**Participants and design.** Participants were 1,058 Americans recruited from Amazon Mechanical Turk (AMT). They were randomly assigned to one of four *gain range* conditions: narrow + exposure, narrow + response, narrow + evaluation, or wide. This sample size was informed by the results of Supplemental Study A, which used a similar paradigm.

**Procedure.** We began by explaining to participants that they would be exposed to a series of lotteries. To reinforce that the chance of winning or losing each lottery was equivalent, we described that each lottery's outcome would be determined by the flip of a coin. Heads would produce a win; tails, a loss. Participants saw one example lottery. This illustrated the format that would be used on the target trials.

On every trial, participants were exposed to the same two number lines—one for gains, one for losses. The gain and loss values for a particular lottery were identified on their respective number line. The gain number line spanned from \$6 to \$32, whereas the loss number line ranged

from -\$20 to -\$6 (see Figure 1). This assured that participants were exposed to the same set of numbers, for Supplemental Study A found that exposure alone was insufficient to place values in the decision sample. Note that 112 unique lotteries—defined by one of 14 different gain values and one of 8 different loss values—can be made from these sets. Although participants in three conditions (wide, narrow + evaluation, narrow + response) saw all 112 lotteries, participants in the narrow + exposure condition only saw 64 of these (i.e., those that could be defined by the 8 gains ranging from \$6 to \$20 and the 8 loss values). Although wide and narrow + exposure participants made accept or reject decisions about every lottery they saw, narrow + evaluation and narrow + response participants were asked a different question about lotteries with higher

### Figure 1

#### *An Example Lottery from Studies 1 and 2*



*Note.* Although this presentation format exposed all participants to a wide range of gains on each trial, whether participants saw lotteries defined by high gains (those from \$22 to \$32) and what judgment or decision participants made on such high-gain trials varied by gain range condition.



gains (\$22 to \$32). In every condition, the trials appeared in random order. The details of these manipulations are described below and summarized in Figure 2:

*Wide.* These participants saw and made decisions about all 112 lotteries. For each lottery, participants indicated whether they would accept or reject it.

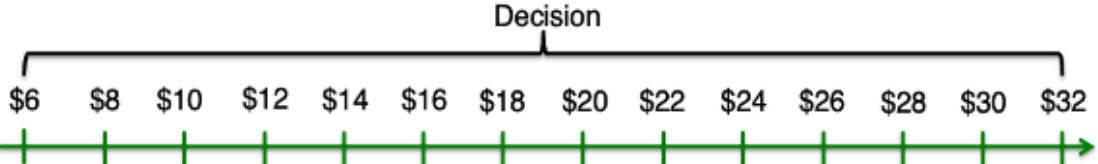
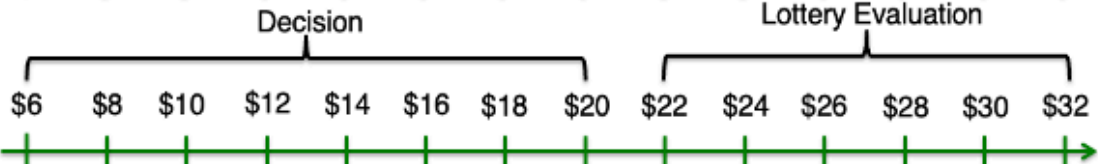
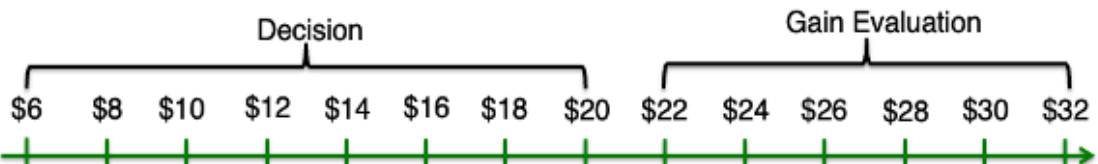
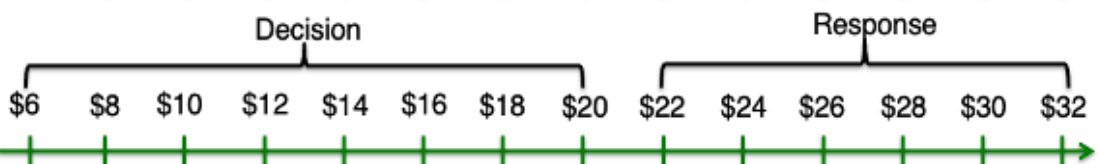
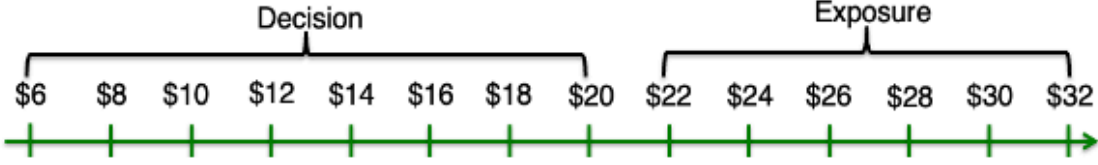
*Narrow + Evaluation.* Whereas participants in this condition also saw all 112 lotteries, they only made a decision about whether to accept or reject lotteries over a narrow gain range: the 64 lotteries that offered the chance to win between \$6 and \$20. For lotteries with gain values between \$22 and \$32, participants made a judgment that would require them merely to subjectively evaluate this value without making a decision about the gamble. More specifically, they rated the attractiveness of the gamble on a non-numeric slider scale anchored at *not at all attractive* and *extremely attractive*. The slider's default value was the midpoint, which was labeled *somewhat attractive*.

*Narrow + Response.* Participants made decisions about the same 64 lotteries that those in the narrow + evaluation condition did. But for the other lotteries—those with gain values between \$22 and \$32—participants responded in light of the values, but not in a way that required them to form a subjective evaluation. Instead, participants merely typed the gain value.

*Narrow + Exposure.* Like in the other two narrow conditions, narrow + exposure participants made decisions about all 64 lotteries with gains between \$6 and \$20. But in this condition, participants were exposed to the gain values from \$22 to \$32 only on the number line itself. That is, participants did not respond in light of the values. So that participants in this condition would confront roughly the same number of trials as participants in the other two conditions, participants saw the set of 64 narrow-gain-range lotteries (those with gains between \$6 and \$20) twice. Given Supplemental Study A found that exposure to values (i.e., \$22 through

**Figure 2**

*Summary of Gain Range Conditions Used in Studies 1 and 2 to Determine What Places Values in the Decision Sample*

Task for Lotteries Containing Specified Gain Values	Condition	Studies
	Wide	1 and 2
	Narrow + Lottery Evaluation	1 and 2
	Narrow + Gain Evaluation	2
	Narrow + Response	1
	Narrow + Exposure	1 and 2

*Note.* Each row describes a condition that builds on the row below it. For example, those in the narrow + gain evaluation condition did not have to merely respond in light of higher-gain (\$22 to \$32) values but had to respond with a subjective evaluation of that gain. All conditions' lotteries had loss values that ranged from \$6 to \$20, in \$2 increments.

\$32 on the gain number line) is insufficient to place them into the decision sample, this condition serves as a baseline against which to test whether responding to, evaluating, and/or deciding based on attribute values places them in the decision sample.

## Results and Discussion

To determine what leads values to enter the decision sample, we tested how our manipulations influenced participants' degree of loss aversion. We started with our preregistered approach by conducting logistic regressions to compute a loss aversion coefficient for each participant. But considering recent discussions regarding how best to conduct such analyses (Andre & de Langhe, 2021; Walasek & Stewart, 2021), we used complementary analytic approaches. We intersperse discussion of these recent back-and-forths into the presentation of our results to make clear how we aim to sidestep these potential problems. This will allow readers to see both the validity of some of the recently articulated concerns as well as how our own conclusions are not threatened by them.

***Logistic regression.*** To permit direct comparisons with Walasek and Stewart's (2015) efforts, we start by following our preregistered analysis plan to take their approach. The initially planned approach: 1) describes the calculation of each participant's loss aversion coefficient, and 2) determines which values are identified as outliers and excluded from the analyses reported below. By this method, participants' loss aversion—the degree to which their choices reflect sensitivity to loss values as opposed to gain values—is not reflected in any individual decision. Instead, it is a property that characterizes a participant's body of decisions as a whole. For each participant, we performed a logistic regression in which we predicted a particular participant's decision to accept (+1) or reject (-1) a lottery as a function of the lottery's gain value and its loss value.

From this regression, we took the beta for the loss value and divided it by the beta for the gain value. This quotient, once multiplied by a negative one, reflects participants' relative sensitivity to losses (vs. gains) when considering risky decisions. Next, we followed Walasek and Stewart's (2015) two-step exclusion criteria. First, we excluded participants whose regression fit deviance scores were among the top 5% ( $n = 52$ ). Second, we removed anyone left whose loss aversion coefficients were negative (thereby displaying an inexplicable preference for lower-expected-value lotteries;  $n = 47$ ). After exclusions ( $n = 99$ ), 959 participants remained. The interested reader can find results with no exclusions or with different exclusion criteria—for this and all studies—in the Supplemental Materials.

We began by comparing our two most-different conditions. As DbS anticipates, those who made decisions over the wide range of gains (\$2 to \$32) displayed more loss aversion (*Median coefficient* = 1.29; 95% bootstrapped CI = [1.17, 1.39]) than narrow + exposure participants (*Median coefficient* = 1.02; 95% bootstrapped CI = [1.00, 1.04]), who made decisions only over the narrow (\$2 to \$20) range of gains,  $Z = 4.58$ ,  $p < .001$ . Wide gain range participants had to make decisions over, subjectively evaluate, and respond in light of additional values (gains ranging from \$22 to \$32). Which process (or processes) pushed those additional attribute values into the decision sample?

To probe this question, we next examined the loss aversion coefficient in the narrow + response condition. If responding in light of values (instead of merely being exposed to them) pushes values into the decision sample, narrow + response participants should show more loss aversion than narrow + exposure conditions. They did not. Instead, narrow + response participants actually showed a non-significant decline in loss aversion (*Median coefficient* = 1.00; 95% bootstrapped CI = [1.00, 1.02]),  $Z = 1.40$ ,  $p = .163$ .






Next, we compared the narrow + evaluation condition to both of the just-reviewed conditions. If values that feed into subjective valuations are placed into the decision sample, then we would expect a greater loss aversion coefficient in this condition. That is what we observed. Evaluation produced a greater loss aversion coefficient (*Median coefficient* = 1.10; 95% bootstrapped CI = [1.05, 1.26]) than both the narrow + response condition,  $Z = 3.84, p < .001$ , as well as the narrow + exposure condition,  $Z = 2.74, p = .006$ . This supports that subjective evaluation *is* sufficient to place values into the decision sample (see Figure 3).

Does actually making decisions in light of a value—an act that also requires exposure, response, and evaluation—further increase the value’s tendency to enter the decision sample? We found that the wide condition produced a marginally higher loss aversion coefficient than the narrow + evaluation condition,  $Z = 1.80, p = .072$ . Although this comparison did not quite reach statistical significance, note that the meaning of this comparison is itself ambiguous. This is because in the wide condition, the loss aversion coefficient was calculated over a different set of lotteries, one that included higher gains (those from \$22 to \$32) and thus higher expected values than in the other three conditions.

This confound is likely problematic (André & de Langhe, 2021; Walasek et al., 2021; see Mellers & Cooke, 1994, for an analogous critique). Alempaki et al. (2019) found that this issue can provide misleading support for DbS in the context of examining loss aversion. First, people may choose to play it safe as the stakes grow (Ferh-Duda et al., 2020; Holt & Laury, 2005). Second, diminishing sensitivity to gains as those gain values grow can also produce what looks like enhanced loss aversion (André & de Langhe, 2021). Both factors will inflate the loss aversion coefficient in the wide condition.

**Figure 3**

*The Characteristics of the Lotteries and Two Indices of Loss Aversion, by Condition (Study 1)*

Condition	Range of Gains for Lottery Decisions	Range of Gains for All Lotteries	Range of Losses	Interval Between Values (Losses, Gains)	Non-Decision Task	Area Under Indifference Curve (AUIC)	Median Loss Aversion Coefficient with 95% Bootstrapped Confidence Intervals
Wide	[\$6, \$32]	[\$6, \$32]	[-\$20, -\$6]	\$2	N/A	N/A	
Wide (Reanalysis, Narrow lotteries)	[\$6, \$20]*	[\$6, \$32]	[-\$20, -\$6]	\$2	N/A	[.34, .36 <sup>a</sup> , .40]	
Narrow + Lottery Evaluation	[\$6, \$20]	[\$6, \$32]	[-\$20, -\$6]	\$2	evaluate attractiveness of lottery	[.35, .39 <sup>a</sup> , .43]	
Narrow + Response	[\$6, \$20]	[\$6, \$32]	[-\$20, -\$6]	\$2	retype gain	[.43, .45 <sup>b</sup> , .47]	
Narrow + Exposure	[\$6, \$20]	[\$6, \$32]	[-\$20, -\$6]	\$2	N/A	[.44, .45 <sup>b</sup> , .48]	

*Note.* For Area Under Indifference Curve (AUIC) calculations, data is presented in the form [lower bound of 95% bootstrapped CI, median AUIC, upper bound of 95% bootstrapped CI]. AUIC values can range from 0 to 1, such that smaller values reflect greater loss aversion. Median AUIC values that do not share a superscript differ at the  $p < .05$  level.

\*Although these participants made decisions about lotteries with gain values in the range [\$6, \$32], only those lotteries with gain values in the range [\$6, \$20] were used to calculate AUIC and the loss aversion coefficient.

Fortunately, we can sidestep these issues by conducting a reanalysis that eliminates the confound. To do so, we recomputed a loss aversion coefficient for those in the wide condition using *only* those lotteries with gains ranging from \$6 to \$20. Just as evaluating high-value gain lotteries influenced narrow + evaluation participants' decisions on the low-value gain lotteries, we can test whether wide lottery participants' actual decisions had the same or perhaps a greater effect. In this way, loss aversion is calculated using an identical set of lotteries in all conditions. With this adjustment, the wide condition no longer produced (even marginally) greater loss aversion (*Median coefficient* = 1.25; 95% bootstrapped CI = [1.12, 1.43]) than the narrow + evaluation condition,  $Z = 1.37$ ,  $p = .172$ . This suggests that actually making decisions did not contribute to placing values in the decision sample; only evaluation had this effect.

***Area under indifference curve (AUC).*** More recently, Walasek and Stewart (2019, 2021) argued that the logistic-regression approach just described—although commonly used in studies of loss aversion—is not a precise instrument for recovering the loss aversion coefficient. Such concerns particularly apply most strongly when estimating large (e.g., > 2.5) loss-aversion coefficients (Walasek & Stewart, 2021), though they still apply when estimating the smaller coefficients observed in the current work. Furthermore, such concerns apply more to the confidence with which a particular individual's loss-aversion parameter can be estimated (given the imprecision of the method for estimating it) but apply less “when making estimates about a group of participants rather than individual participants” (Walasek & Stewart, 2021, p. 11), as we do in the current work.<sup>2</sup> That said, Walasek and Stewart (2021) recommend an alternative way to

---

<sup>2</sup> Notably, André and de Langhe (2021) were unable to replicate Walasek and Stewart's (2015) findings when reanalyzing the lotteries that were common to each condition using the logistic-regression approach. In contrast, our Study 1 (and Study 2) shows DbS-consistent between-condition differences using even this (imperfect) approach. We suspect the difference can be explained by two factors: 1) the present studies recruited much larger samples than does Walasek and Stewart (2015), and 2) the reanalysis of Walasek and Stewart had to be done over only 9 lotteries per participant whereas there were 64 lotteries that were common across the conditions in the present studies. As a result, the power offered by the present studies was presumably able to overcome the imprecision of the measure.



capture loss-averse behavior in these paradigms, what they call Area Under the Indifference Curve (AUC). As reported in the Supplemental Materials (and reported in Figure 3), analyses using the AUC approach replicated those using the preregistered logistic regression approach.

***An artifact of participant disengagement?*** In each condition, participants were exposed to more than 100 mixed gambles and had to complete some action (i.e., retyping, evaluation, and/or making a decision) for each. Such repetition has the potential to produce participant fatigue and disengagement. If some participants responded to such repetition with a shift to random responding, it becomes quite relevant to consider whether that would be sufficient to produce the patterns of results we observed. Fortunately, three reasons suggest this possibility does not pose a threat to the study's internal validity. First, our preregistered analysis plan was designed to flag and exclude inattentive respondents (e.g., outliers, those whose decisions were not sensitive to lotteries' expected values). Second, Walasek and Stewart (2015) ran a simulation showing that if participants adopted a simplifying rule like accepting half of the gambles (or the half of gambles with the highest expected values), then this has an influence on the logistic regressions' intercept, but not the slopes, which are used to calculate our key dependent measure, loss aversion. Third, our predictions anticipated between-condition *differences* in loss aversion calculated over the exact same set of lotteries (i.e., those with gain values between \$6 and \$20), instead of loss aversion coefficients of a *specific* value (that could themselves be a function of some participant fatigue). Furthermore, and most convincingly addressing the concern that disengagement would produce the between-condition differences we observed, participants in all three "narrow +" conditions made decisions about the exact same set of lotteries. For this reason, even if participant fatigue did have an influence on responding and the indexes of loss aversion, then this would have biased between-condition comparisons toward the null.

## Study 2

Although Study 1 identified subjective evaluation as the crucial process that places attribute values into a decision sample, Study 2 answers the question of precisely *what* must be evaluated. In Study 1, participants offered subjective valuations of the risky prospect as a whole, which included a possible gain *and* a possible loss. Our assumption was that evaluating the lottery requires evaluation of its components, the loss *and* the gain. But if our mechanistic reasoning is correct, then it should be sufficient for participants to evaluate *only* the gain in order to produce the same effect. Such a manipulation would offer a more conservative test of the evaluation account, testing whether subjectively evaluating additional potential gains on their own would influence how other gains were then interpreted and relied upon as a component of a mixed gamble.

Study 2 retains our two baseline conditions: narrow + exposure and wide. In addition to the narrow + (lottery) evaluation condition used in Study 1, we added a narrow + gain evaluation condition. If our mechanistic account is correct—if values enter decision samples because the values themselves (as opposed to the prospects they help to define) are evaluated—then we should find that both the narrow + lottery evaluation and the narrow + gain evaluation conditions produce greater loss aversion than the narrow + exposure condition. Furthermore, we should expect these two evaluation conditions to have similar effects. If instead the narrow + lottery evaluation is unique in elevating loss aversion, this would suggest a more nuanced role of evaluation in placing values in the decision sample, one that would require us to revise our evaluation account.

## Method

**Participants and design.** Participants were 2,223 Americans recruited via AMT.

Participants were randomly assigned to one of four gain range conditions: narrow + exposure, narrow + gain evaluation, narrow + lottery evaluation, or wide.

**Procedure.** The procedure was similar to the one used in Study 1. The narrow + exposure, narrow + lottery evaluation, and wide conditions were all repeated from the previous study. The difference between the new narrow + gain evaluation and the previously used narrow + lottery evaluation condition is *what* participants subjectively evaluated. In both conditions, participants offered evaluations related to lotteries whose gains ranged from \$22 to \$32. Whereas those in the narrow + lottery evaluation condition evaluated the attractiveness of the lottery, those in the new narrow + gain evaluation condition evaluated the attractiveness of the possible gain. These ratings were collected on the same unnumbered slider scale from Study 1, anchored at *not at all* and *extremely attractive*. The midpoint—where, by default, the slider began—was again labeled *somewhat attractive*.

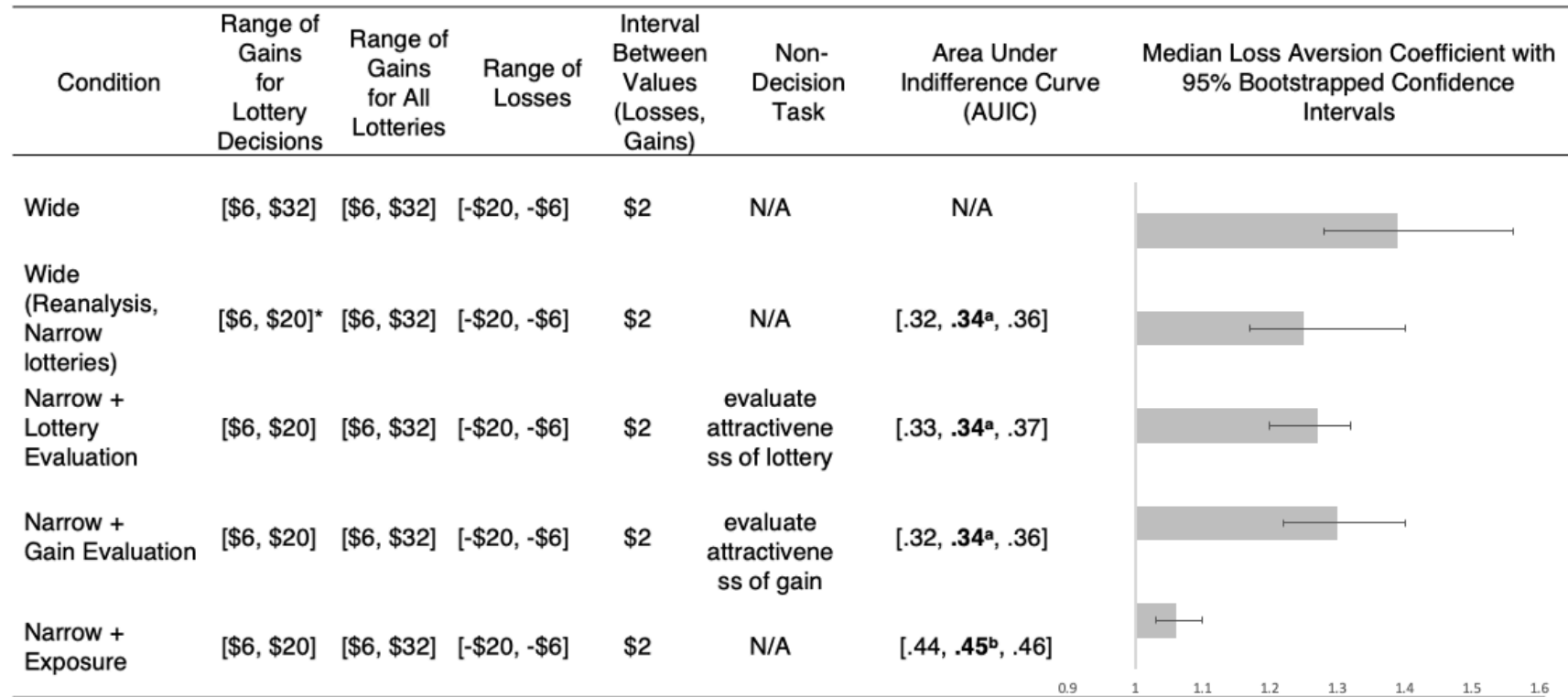
## Results and Discussion

**Logistic regression.** We calculated loss aversion coefficients and trimmed the data using the same preregistered procedures as described for Study 1. First, we eliminated anyone whose regression fit deviance scores were among the top 5% ( $n=111$ ). Second, we eliminated anyone left who had a negative loss aversion coefficient ( $n=106$ ). In total, 217 participants were excluded based on Walasek and Stewart's (2015) exclusion criteria. This left 2,006 participants for the final analysis. The Supplemental Materials include analyses applying both more stringent and laxer (i.e., no) exclusion criteria.

As displayed in Figure 4, we again found that participants displayed greater loss aversion when they made decisions over a wide range of gains (*Median coefficient* = 1.39; bootstrapped

**Figure 4**

*The Characteristics of the Lotteries and Two Indices of Loss Aversion, by Condition (Study 2)*



*Note.* For Area Under Indifference Curve (AUC) calculations, data is presented in the form [lower bound of 95% bootstrapped CI, median AUC, upper bound of 95% bootstrapped CI]. AUC values can range from 0 to 1, such that smaller values reflect greater loss aversion. Median AUC values that do not share a superscript differ at the  $p < .05$  level. \*Although all these participants made decisions about lotteries with gain values in the range [\$6, \$32], only those lotteries with gain values in the range [\$6, \$20] were used to calculate AUC and the loss aversion coefficient.

95% bootstrapped CI = [1.28, 1.56]) compared to those (narrow + exposure participants) who were merely exposed to the values at the upper end of that range (*Median coefficient* = 1.06; 95% bootstrapped CI = [1.03, 1.10]),  $Z = 7.60, p < .001$ . Furthermore, we replicated the Study 1 result that evaluating the attractiveness of the wider range of lotteries (narrow + lottery evaluation) also elevated loss aversion (*Median coefficient* = 1.27; 95% bootstrapped CI = [1.20, 1.32]) compared to the narrow + exposure gain range condition,  $Z = 5.15, p < .001$ .

Participants in the new narrow + gain evaluation condition showed elevated loss aversion (*Median coefficient* = 1.30; bootstrapped 95% CI = [1.22, 1.40]) compared to the narrow + exposure condition,  $Z = 5.82, p < .001$ . Furthermore, the two evaluation conditions were not statistically distinguishable,  $Z < 1$ . In other words, evaluation—whether of the lottery (meaning the gain and loss together) or the gain value directly—was sufficient to place gain values in the decision sample.

Finally, we considered the role of actually making decisions based on the values (as opposed to merely evaluating them) by comparing the two evaluation conditions to the wide gain range condition. Those in the wide condition showed marginally greater loss aversion than those in the narrow + gain evaluation condition,  $Z = 1.84, p = .066$ , and significantly more than those in the narrow + lottery evaluation,  $Z = 2.43, p = .015$ . But recall the concern raised in Study 1: Did this elevated loss aversion reflect the influence of making a decision on placing the gain values from \$22 to \$32 in the decision sample, or did it reflect that wide condition participants' loss aversion coefficient was calculated over lotteries that included those gain values? To disentangle these possibilities, we recalculated the wide condition's loss aversion coefficients using only those lotteries that all participants accepted or rejected—i.e., those with gain values ranging from \$6 to \$20. This reduced the loss aversion observed in the wide condition (*Median*

*coefficient* = 1.25; 95% bootstrapped CI = [1.17, 1.40]) so that it was no longer greater than that observed in the narrow + lottery evaluation,  $Z = 0.75$ ,  $p = .453$ , or the narrow + gain evaluation conditions,  $Z = 1.24$ ,  $p = .217$ . Once again, we found that it was only evaluation that was responsible for placing values in the decision sample.

**AUIC.** Like in Study 1, we reconducted these analyses using Walasek and Stewart's (2021) AUIC method. As reported in the Supplemental Materials (and described in Figure 4), these results provided continued support that it is the evaluation of values (and not merely the fuller prospects those values help to define) that is responsible for placing those values in the decision sample.

### Study 3

Study 3 extended our investigation to a new domain: patience. Participants considered tradeoffs between smaller-sooner and larger-later monetary payouts. Previous research has found that people express more impatient preferences for smaller-sooner rewards when they feel less connected to their future selves (Ersner-Hershfield et al., 2009; Hershfield & Bartels, 2018; cf. Frederick, 2003), after being induced to feel financially deprived (Callan et al., 2011), after completing a visceral need state induction (e.g., hunger; Skrynka & Vincent, 2019), or even after recalling a hasty service experience (e.g., a fast-food restaurant visit; DeVoe et al., 2013). But much as we showed that loss aversion can vary based on the set of attribute values people have recently evaluated, we considered how DbS would expect patience (i.e., delay discounting) to vary for similar reasons. More specifically, we varied the set of temporal values (instead of monetary values, as in Studies 1-2) that might enter participants' decision samples and affect their willingness to delay payoffs. After all, people's willingness to accept delays should depend on how subjectively short they seem.

Although all participants made choices over the same set of tradeoffs, they either subjectively evaluated or merely responded to (i.e., retyped) other temporal values that came from a uniform or skewed (with many near times and few distal times) distribution. If evaluation introduces these values into the decision sample, then the nature of the distribution should have a stronger influence on evaluation-condition participants' display of patience in a way that DbS would expect. More specifically, those who evaluate (as opposed to merely retype) values from the uniform distribution should display more patience than those considering the skewed distribution. This is because in the context of having considered and evaluated many objectively short delays (as those in the skewed time distribution condition do), even moderate delays should seem like long waits.

## Method

**Participants and design.** We recruited 1,209 Americans from AMT. Participants were randomly assigned to one of four conditions in a 2(time distribution: uniform or skewed) X 2(task: response or evaluation) full-factorial design. All participants are included in the analyses reported below. In the Supplemental Materials, we report results using a more stringent inclusion criterion.

**Procedure.** Participants learned they would see 60 pairs of payoffs. Each pair would feature a certain amount of money that could be received immediately or a larger amount of money that would be received after a certain time delay. The time distribution manipulation determined whether those time delays came from a highly skewed distribution (1 day, 1 week, 1 month, 2 months, 6 months, 12 months) or a relatively uniform distribution (2 months, 4 months, 6 months, 9 months, 12 months, 15 months). As displayed in Figure 5, three of the time delays are common to the conditions: 2 months, 6 months, and 12 months. Crucially, the ranks of these

values in each condition were either highly discrepant (2 months: 3<sup>rd</sup> vs. 6<sup>th</sup>), somewhat discrepant (6 months: 2<sup>nd</sup> vs. 4<sup>th</sup>), or barely discrepant (12 months: 1<sup>st</sup> vs. 2<sup>nd</sup>). Or considered differently, the three delays are more similar in rank in the skewed distribution (occupying the adjacent 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> positions), whereas these same delays have more differentiated ranks in the relatively uniform distribution (occupying the non-adjacent 2<sup>nd</sup>, 4<sup>th</sup>, and 6<sup>th</sup> positions).

The immediate payoffs were \$100, \$200, \$300, or \$400. The delayed payoffs were \$200, \$300, \$400, or \$500. We generated tradeoffs using only those 10 combinations for which the delayed payoff would be greater than the immediate payoff. And because in each condition the time delay for the larger payoff could take one of 6 distinct forms, this means each participant saw 60 tradeoffs. Thirty of the pairs were common to both conditions. (These were the ones that involved delays of 2, 6, or 12 months.) On these choice trials, participants indicated whether they would prefer the smaller amount today or the larger amount after the specified delay. But instead for the other thirty pairs—those that involved delays that were unique to one of the time distribution conditions—participants did not indicate their preferred choice. Instead, what they did depended on their task condition. In the *response* task condition, participants were asked to “type the amount of time you will have to wait to receive the larger payout.” In the *evaluation* task condition, participants were asked to indicate “how unappealing it would be to have to wait [time delay] for the payoff.” Participants responded on a slider scale that ranged from “*not at all unappealing*” to “*extremely unappealing*.” The middle was labeled “*somewhat unappealing*.”

To make sure that responding to (i.e., retyping) or evaluating values had a chance to modify participants’ decision samples before their very first choice trials, we had all participants consider these three time delays—i.e., those specific to each condition and that would be used on the subsequent response or evaluation trials—before beginning the main task. Response



**Figure 5**

*Time Delays Seen by Those in the (A) Skewed and (B) Uniform Time Distribution Conditions (Study 3)*

**(A) Skewed Time Distribution**

1 day 1 week 1 month



**(B) Uniform Time Distribution**



*Note.* Time delays above each timeline are unique to that condition and are those time delays participants responded to (by retyping them) or subjectively evaluated, depending on their Task condition. Time delays below the timeline are common to both time distribution conditions and are those delays used to measure patience. The values in parentheses indicate what proportion of the other time delays shown in that condition are longer than that time delay. If all values on a timeline do indeed compose the decision sample for participants in that condition, decision by sampling anticipates less patience in the skewed time distribution condition (because the parenthetical values are lower there) and less sensitivity to the shift from 2 to 12 months (given  $2/5 - 0/5 < 5/5 - 1/5$ ).

participants had to retype the three time delays. Evaluation participants had to indicate how unappealing it would be to wait those amounts of time for a reward of \$200 to \$500. At that point, the 60 tradeoffs appeared in random order.

## Results and Discussion

We wanted to understand how our manipulations—of both time distribution and task—affected participants' patience. Participants could display patience (opting for the larger-later reward) or impatience (choosing the smaller-sooner reward) on each trial (as opposed to through their revealed sensitivity to changes in loss and gain values across their choices). We used mixed models to explain variation in participants' responses to each trial. All models include two fixed effects—the monetary values of the shorter-sooner and the larger-later reward—that simply serve as covariates. Each model also includes a random effect of participant. This accounts for the non-independence of each participant's 30 choices. Although concerns that R's lme4 package inflate Type 1 error—from 5% to roughly 8%—primarily apply to much smaller samples than what the present studies included, we were mindful of this issue when deciding to use SPSS's MIXED function, whose Satterthwaite's degrees of freedom approximation is largely immune to these concerns (Luke, 2017).

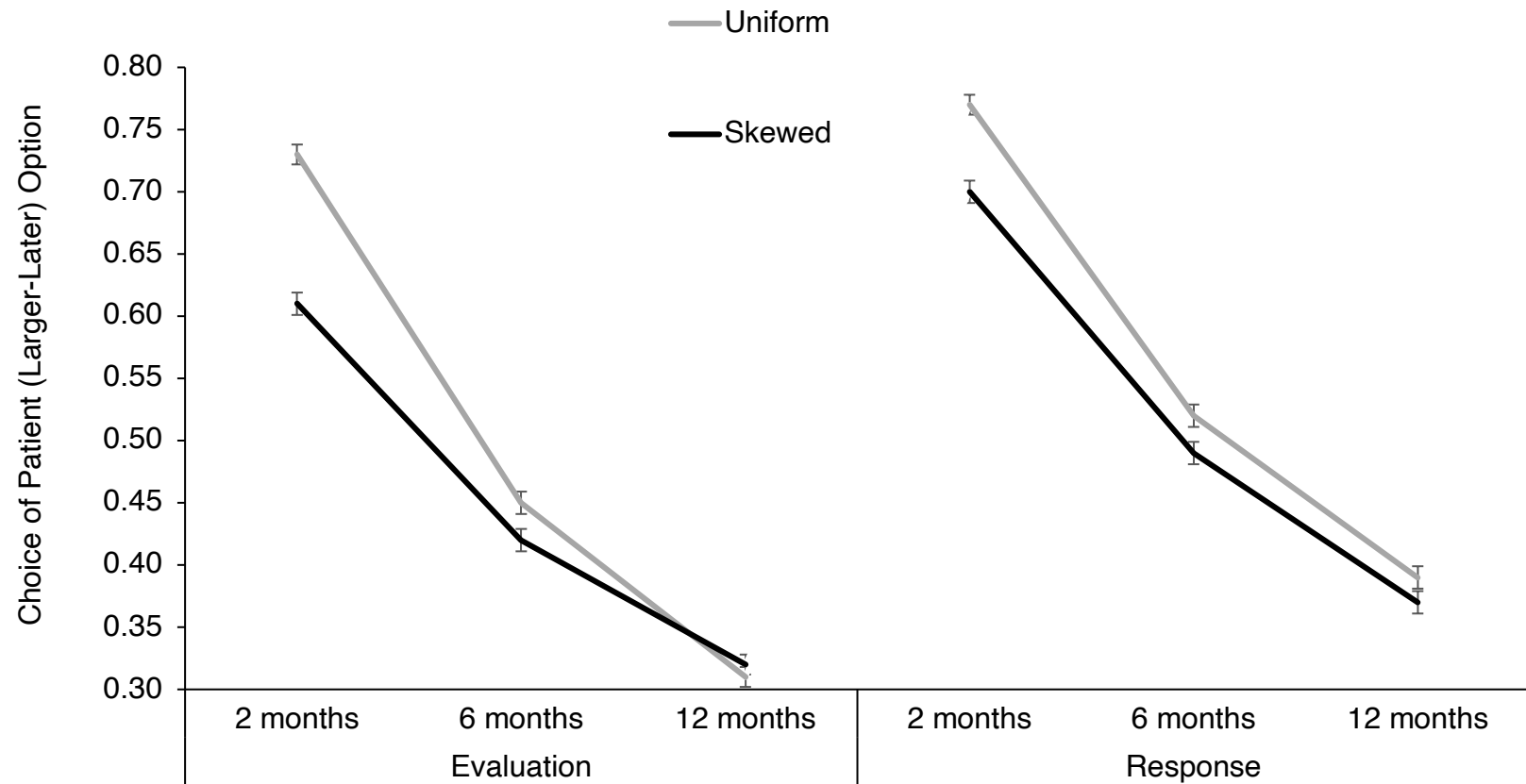
We began by testing whether those exposed to the uniform distribution of time delays displayed more patience than those who saw the skewed distribution. After all, the required delays for the larger reward should have seemed subjectively shorter when considered in the context of the uniform than the skewed distribution. And indeed, this first prediction was confirmed. When considering the same tradeoff, uniform participants indicated a willingness to wait for the larger reward on more trials (52.61%) than did skewed participants (48.42%),  $t(1205.97) = 2.28, p = .023$ .

Did such patience take the form that DbS would anticipate? More specifically, we expected that uniform participants' greater patience would emerge most clearly at 2 months and least strongly at 12 months. This is because the ranking of this delay was maximally discrepant between the two distribution conditions in the former case and the least discrepant in the latter case. We coded *time delay* as -1 (2 months), 0 (6 months), and +1 (12 months). As expected, we observed a Time Distribution X Time Delay interaction,  $t(35057.99) = 12.04, p < .001$ . This interaction reflected the expected pattern. Uniform (compared to skewed) participants were much more patient at 2 months,  $t = 4.87, p < .001$ ; somewhat more patient at 6 months,  $t = 2.28, p = .023$ ; and no more patient than skewed participants at 12 months,  $t < 1$ .

To address our central question—whether it is the evaluation that places values in the decision sample—we tested whether the just-reported findings were driven by participants in the evaluation (compared to the response) condition. Supporting this hypothesis, the Time Distribution X Time Delay X Task interaction was significant,  $t(35055.99) = 4.85, p < .001$ . When participants subjectively evaluated the additional values (those that caused the overall distribution to be skewed or uniform), a significant Time Distribution X Time Delay interaction gave clear evidence that such values populated the decision sample and influenced patience as decision by sampling would predict,  $t(35055.99) = 12.08, p < .001$ . In contrast, when participants merely retyped the additional values, this Time Distribution X Time Delay interaction was much weaker,  $t(35055.99) = 4.78, p < .001$  (see Figure 6). Whereas Studies 1 and 2 showed that evaluation, not simply responding in light of the values (or mere exposure), was necessary to place values in the decision sample, Study 3 did find that responding in light of the values was sufficient. That said, being prompted to evaluate the values led to much stronger effects. One speculative possibility is that even when Study 3 participants merely retyped the values, some of

**Figure 6**

The Proportion of Participant Choices Reflecting Patience as a Function of Task Condition, Time Distribution, and Delay



*Note.* Error bars reflect  $\pm 1$  standard error from the mean. (Study 3).

them may have spontaneously evaluated the durations.

Study 3 extends our investigation to a new domain, patience. Through this change, the study provides convergent evidence that evaluation places values in the decision sample. But the study also included a new feature that did not characterize our first two studies or previous studies that used the loss-aversion paradigm (e.g., Walasek & Stewart, 2015). More specifically, we introduced our key manipulation—evaluating or responding to those time delays that were unique to each time distribution conditions—*before* participants responded to any of the patience measures. In this way, our manipulations had the potential to influence participants' responses from the very first trial. In contrast, our manipulations in Studies 1 and 2 could not have had an effect until they were at least a few decisions into their more than 60 decisions.

With this methodological revision in mind, we returned to our investigation of loss aversion. But this time, we included the design modification introduced in Study 3. In a preregistered study (Supplemental Study B,  $N = 900$  Americans from Mechanical Turk), we had all participants either respond to (retype) or evaluate those gain values that constituted our manipulation (those from \$22 to \$32) *before* completing any actual lottery decisions. Otherwise, the procedure followed that of the narrow + response condition from Study 1 and the narrow + gain evaluation condition from Study 2. As expected, we replicated our key result: Those in the narrow + gain evaluation condition displayed greater loss aversion (*Median coefficient* = 1.19, 95% bootstrapped CI = [1.12, 1.26]) than did those in the narrow + response condition (*Median coefficient* = 1.02, 95% bootstrapped CI = [1.00, 1.05]),  $Z = 4.06$ ,  $p < .001$ . The AUIC approach yielded similar findings: We saw greater evidence of loss-averse behavior in the evaluation (*Median* = .33) compared to the response condition (*Median* = .45),  $Z = 6.22$ ,  $p < .001$ .

In analyses of a new memory measure, we found those in the evaluation condition did not have a superior memory for the attribute values compared to those in the response condition,  $t < 1$ . Furthermore, superior memory of the gain values did not significantly correlate with either index of loss aversion: the loss aversion coefficient calculated using the logistic-regression approach,  $r = -.02$ , or the AUIC,  $r = -.06$ . These null effects buttress similar findings from Walasek and Stewart (2019) that attribute values' existence in the decision sample is independent of their memorability (see Supplemental Materials). To summarize, subjectively evaluating values—whether such evaluations were only interspersed with (Studies 1-2) or also occurred in advance of key trials (Studies 3, B)—seemed to place those values in the decision sample and influence decision making just as DbS anticipates.

#### Study 4

Study 4 built on our previous studies in two ways. First, we moved to a new decision context and a new paradigm for testing whether evaluation places attribute values in the decision sample. Participants considered whether they would be interested in receiving a vaccine for a novel disease. Even before the current COVID-19 pandemic, psychologists have been aware of the problem of—and challenges of addressing—vaccine hesitancy (e.g., Rossen et al., 2016). Vaccine hesitancy has been connected to individual differences like social or political identity as well as conspiratorial thinking and distrust in traditional medicine (Hornsey et al., 2020; see Hornsey et al., 2021). But much as loss aversion and patience were influenced by values placed into the decision sample through evaluation, it seemed possible that vaccine hesitancy could serve as yet another domain in which to test for analogous effects.

In Study 4, participants repeatedly indicated their preference for various versions of a vaccine defined by different combinations of vaccine efficacy (5% to 95%) and side-effect

duration (1 to 10 days of flu-like symptoms). We randomly assigned participants to make decision about the vaccines (i.e., indicate whether they would accept or reject that version of the vaccine) that were high efficacy (55% to 95%) or low efficacy (5% to 45%). For the other vaccines—those about which participants did not make a decision (i.e., those in the alternative efficacy range)—participants either retyped the efficacy value (response condition) or subjectively evaluated the attractiveness of the efficacy value (evaluation condition). We expected to conceptually replicate our earlier results that evaluation places values in the decision sample. This would be reflected by the willingness to accept the high-efficacy vaccines (compared to the low-efficacy vaccines) being magnified in the evaluation (compared to the response) condition.

Second, we wanted to probe more directly why it was that evaluating (as opposed to retyping) certain efficacy values may change participants' interest in vaccines that are defined by other values. Recall that both Supplemental Study B as well as previous research (Walasek & Stewart, 2019) found that memory probes do not recreate the decision sample that guides subsequent decision making. Why is this? One possibility is that these memory probes—as explicit requests to scour one's memory stores for recently encountered attribute values—may not be a valid probe of which values, unprompted, are naturally accessible in working memory and thus part of the decision sample. Though there is an alternative possibility. Perhaps the decision sample is not itself recruited at the time of decision making, but instead merely describes the set of values that have previously served to shape people's sense of what values are relatively large or small.

That is, people may possess a fluid, malleable, subjective numeric evaluation scale that they use to subjectively characterize newly encountered values. Such flexible standards can of

course be domain specific, shaping one's sense of what makes, for example, a good salary or a tolerable time delay. But crucially, this scale may be constructed and updated just as DbS anticipates, with the rank ordering of previously encountered values helping to define what constitutes a relatively small, medium, or large value. Furthermore, this may be why subjectively evaluating newly encountered values—thereby forcing one to refine one's sense of what values are large vs. small, appealing vs. not—is what seems to place such values in the decision sample. By this understanding, evaluated values may not be recruited online to make sense of a newly encountered value (thus explaining why memory probes do not reproduce their apparent prominence in guiding new evaluations). Instead, it may be the remnants of *evaluating* such values—thereby encouraging updates to one's internal subjective evaluation scale—that guide the interpretation of new values.

Although this theoretical account speculatively characterizes our previous results, Study 4 offers an initial test of this process by aiming to directly probe this subjective evaluation scale that previously evaluated values may help to shape. Toward this end, after participants completed all the vaccine trials, we asked them to subjectively evaluate the five efficacy levels about which they had made vaccine decisions. We expected that participants' subjective evaluation of these values would be more polarized (i.e., the ratings of the high versus low efficacy values would be more differentiated) when participants had subjectively evaluated (as opposed to merely responded to) the other efficacy attribute values. This would reflect that actually evaluating (instead of merely responding to) other values was shaping one's subjective evaluation scale for what constituted a high versus low value. Finally, we expected that these subjective evaluations would mediate the central effect described earlier. That is, we predicted that evaluation-task participants' more extreme evaluations of the efficacy levels would explain



their more polarized patterns of decision making. If so, this would more directly capture the process by which evaluating values changes decision making regarding newly encountered values.

## Method

**Participants and design.** We recruited 1,224 Americans from AMT. Participants were randomly assigned to one of four conditions in a 2(decision range: low or high) X 2(task: response or evaluation) full-factorial design. Once again, all participants are included in the analyses reported below. In the Supplemental Materials, we report results using a more stringent inclusion criterion.

**Procedure.** Before beginning the main study, participants completed three items designed to measure baseline *vaccine interest*: “In general (i.e., not simply with regard to the COVID-19 vaccines), are you more *interested in* or *skeptical about* being vaccinated?” (1 = extremely skeptical, 10 = extremely interested in), “In the last 10 years, how many of those years have you received the flu vaccine?”<sup>3</sup>, and “In general, how do you feel about the COVID-19 vaccines being used in the United States?” (1 = extremely negative, 10 = extremely positively). The items displayed good internal reliability ( $\alpha = .71$ ) and thus were averaged to create a vaccine interest composite ( $M = 6.90$ ,  $SD = 2.18$ ).

At that point, participants were told they would be asked to consider a fictitious novel infectious disease, Minerva X-35:

“The average person who gets Minerva X-35 experiences about a week of severe flu-like symptoms: fever, cough, difficulty breathing, fatigue, and head and body aches.

---

<sup>3</sup> Due to a programming mistake, participants had to answer between 1 and 10. This item should have asked about flu vaccine uptake in the last 9 years and offered responses between 0 and 9. Regardless, the item correlated with the other two items. By excluding this item from the three-item composite, all key effects reported below directionally strengthen. To err toward conservatism, we retain the item in the composite.

Approximately 7% of people who get Minerva X-35 require hospitalization. Of those who are hospitalized, 5 to 6% die. Unlike with COVID-19, the infected's likelihood of hospitalization or death does not depend on age."

Participants then read the following information about a vaccine called Zenoa:

"Due to the genetic instability of the virus, there is much uncertainty about how effective any vaccine that is developed will ultimately be. As a result, governmental agencies are interested in learning about how people would evaluate different possible vaccines that might emerge. As a result, you will consider vaccines that vary in terms of their efficacy (i.e., the reduction in the likelihood of infection) and side effects (i.e., the number of days you are likely to be bedridden with flu-like symptoms)."

Ultimately, all participants saw the same 100 versions of the Zenoa vaccine that varied along two dimensions: efficacy (5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, 95%) and side-effect duration (in days: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10). The 100 trials appeared in a random order.

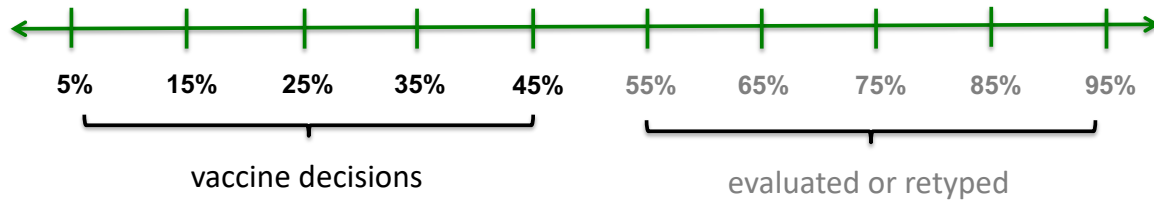
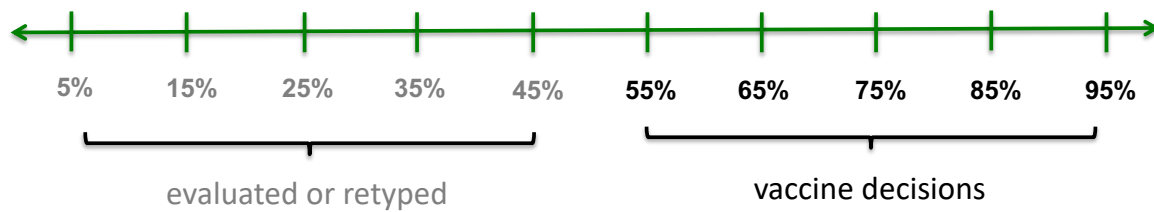
**Decision range manipulation.** For half of the trials, participants made a decision about whether they would elect to take Zenoa if it ended up having the specified properties (efficacy, side-effect duration). For participants in the *high* decision range condition, they made this decision whenever considering a vaccine with relatively high efficacy (55% or higher). For participants in the *low* decision range condition, they made the decision whenever considering a vaccine with relatively low efficacy (45% or lower).

**Task manipulation.** On the 50 non-decision trials, those in the *response* task condition were asked to retype the efficacy of the vaccine they were considering. In this way, their attention was drawn to the number, and participants had to supply response in light of it. Instead, those in the *evaluation* task condition were asked how appealing a particular version of the

Zenoa vaccine was in terms of its stated efficacy. Participants responded on a non-numeric slider scale that was anchored at “not at all appealing” and “extremely appealing”, with the middle of the scale labeled “somewhat appealing.” In this way, these participants’ attention was also drawn to the number, though their response went beyond simply retyping it to require a subjective evaluation of the attribute value. The procedure is summarized in Figure 7.

**Pre-trials intervention.** Given all 100 trials appeared in a random order, the decision and task trials were interspersed. As in Study 3 (and Supplemental Study B), we had all participants complete a condensed version of the task manipulation *before* any of these 100 trials began. This allowed the key task manipulation to influence decisions from the first trial. All participants were exposed to the five efficacy levels that would ultimately characterize the vaccines on the task trials. Response participants had to retype each value (just as they would on the task trials). Evaluation participants had to subjectively evaluate the attractiveness of each efficacy level (just as they would on the task trials). The only difference with the task trials was that a specific side-effect level was not paired with each efficacy level. Instead, the instructions noted that the vaccine would bring “with it between 1 and 10 days of side effects (with flu-like symptoms.)”

**Post-trials evaluation measure.** After completing the 100 trials (50 decisions, 50 tasks), participants were asked to subjectively evaluate the 5 efficacy levels that were used in the decision trials. That is, low decision range participants evaluated 5%, 15%, 25%, 35%, and 45% efficacy. High decision range participants evaluated 55%, 65%, 75%, 85%, and 95% efficacy. More specifically, participants responded to the prompt, “Please rate how appealing it would be if Zenoa’s efficacy at preventing Minerva X-35 were each of the following.” Participants moved a slider from the left end of a 101-point scale to their final response. The actual numbers (0 to 100) corresponding to each slider position were not visible to participants (to keep them from

**Figure 7***Study 4 Materials by Decision Range Condition***(A) Low Decision Range****(B) High Decision Range**

*Note.* Each vaccine efficacy level was paired, across different trials, with all 10 side-effect duration levels (5%, 15%, 25%, 35%, 45%, 55%, 65%, 75%, 85%, 95%).

simply responding with the efficacy percentage.) Instead, participants (like on the evaluation task trials) saw a left-anchor (“not at all appealing”), a mid-point label (“somewhat appealing”), and a right-anchor (“extremely appealing”).

**Results and Discussion**

We began by testing whether there was evidence that the evaluation task (more than the response task) placed those task values into the decision sample, with predictable influence on vaccine decisions. This multi-level model included fixed effects of decision range (+1: high, -1: low) and task (+1: evaluation, -1: response) that characterized each participant’s condition assignment. Crucially, the interaction between these two variables was included as well.

Furthermore, we included fixed effects of side effect and efficacy that described the vaccine's properties on a particular decision trial. These predictors merely served as covariates. The side-effect duration was centered by rescaling the values (1 to 10 days) to range from -4.5 to +4.5. The efficacy values were recoded by decision range condition to describe whether a particular decision trial's vaccine was relatively efficacious for the decision trials that *that* participant confronted: -2 (5% or 55%), -1 (15% or 65%), 0 (25% or 75%), 1 (35% or 85%), +2 (45% or 95%). To account for nonindependence across trials, we included participant as a random factor.

Unsurprisingly, we observed main effects of baseline vaccine support,  $t(1219) = 15.37, p < .001$ , decision range,  $t(1219) = 15.52, p < .001$ , efficacy,  $t(59974) = 63.46, p < .001$ , and side-effect duration,  $t(59974) = -45.49, p < .001$ . In other words, people expressed more willingness to get a vaccine when they entered the experiment with more support for vaccines, they made decisions about more efficacious vaccines, they considered a vaccine that was more efficacious than the other vaccines they made decisions about, and the vaccine promised fewer days of side effects. Though of central relevance, we observed a Decision Range X Task interaction,  $t(1219) = 2.63, p = .009$ . That we observed no main effect of task,  $t < 1$ , speaks to how the task manipulation operated symmetrically across the two decision ranges.

More specifically, when participants evaluated the other efficacy levels, then participants' stated intention to receive the vaccine strongly depended on the decision range—i.e., whether they were making decisions about a set of quite efficacious vaccines (55% to 95% efficacy) or less efficacious vaccines (5% to 45% efficacy): 68.27% vs. 34.18%. But when participants merely responded to (by retyping) the other efficacy levels, this gap was significantly reduced: 62.37% vs. 38.06%. By the evaluation account, evaluating (more than merely responding to) values places them in the decision sample, meaning they are used in the pool of values to

determine a considered efficacy's subjective magnitude. Evidence of this is seen in the greater spread between vaccine interest in the high and low decision range conditions among those who evaluated (34.09 percentage point difference) versus merely retyped (24.31 percentage point difference) the other attribute values. Our subsequent analyses aim to capture this process more directly.

Next, we moved to our test of whether the evaluation task (compared to the response task) changed how participants subjectively evaluated the efficacy levels about which they made decisions—i.e., the post-trials evaluation measure. Given participants made these judgments about 5 levels of efficacy (i.e., those that composed the vaccines about which they made decisions), we retained efficacy as a predictor but dropped side-effect duration from the model (given such durations were not manipulated on the post-trials evaluation measures). Once again, we observed main effects of vaccine support,  $t(1219) = 15.31, p < .001$ , decision range,  $t(1219) = 17.34, p < .001$ , and efficacy,  $t(4895) = 60.75, p < .001$ . In other words, vaccine efficacies were rated as more attractive by those who entered the experiment with more vaccine support, those who made decisions about (and thus evaluated) more efficacious vaccines, and when a vaccine efficacy level compared favorably to the other efficacies about which decisions were made. There was no main effect of task,  $t < 1$ .

But of central relevance, we again observed a Decision Range X Task interaction,  $t(1219) = 2.09, p = .037$ . This supported our explanation for why the evaluation task manipulation influenced participants' decisions about which vaccines to accept. That is, when participants had evaluated the other efficacy rates (i.e., those about which they did not make vaccine decisions), participants showed a sizable gap in their evaluation of the high versus low range of efficacies:

62.61 vs. 36.26 ( $M_{\text{dif}} = 26.35$ ). But when participants had merely retyped the other efficacy rates, this post-trial evaluation gap shrank by 21%: 60.18 vs. 39.43 ( $M_{\text{dif}} = 20.75$ ).

Finally, we tested whether these (post-trials) efficacy evaluations statistically mediated the previously reported interaction on the vaccine decisions. Toward that end, we added an *evaluation* variable to the original model. This value corresponded to how participants ultimately rated (on the post-trials evaluation measure) the specific efficacy level that described a particular vaccine trial. The proposed mediator (i.e., evaluation) significantly predicted participants' likelihood of accepting the vaccine on a specific trial,  $t(51440.49) = 61.66, p < .001$ . The Decision Range X Task interaction remained significant,  $t(1196.09) = 2.28, p = .023$ . This pattern of results is thus consistent with partial mediation, Sobel  $z = 2.09, p = .037$ .

Whereas in Studies 1 through 3, evaluating attribute values influenced loss aversion and patience presumably because such evaluations changed the way that other attribute values were subjectively evaluated, Study 4 documented this process more directly. It seems that subjectively evaluating values, more than merely needing to respond in light of them (e.g., by retyping them), changed how other values were subjectively evaluated. Although this finding is of basic scientific import in directly documenting our proposed process, it also has practical relevance. In an effort to encourage certain behaviors, it may be more effective to frame certain values as relatively larger or small not merely by offering reference points that serve this goal, but by nudging people to subjectively evaluate those reference points as well.

### **General Discussion**

For decades, behavioral scientists have struggled to understand how people characterize the magnitude of quantitative attributes. DbS offers a relatively new answer. It moves beyond a mere description of the relationship between objective quantities and subjective valuations by

positing that such valuations are arrived at through a comparative process. The theory posits that people subjectively assess attribute values by comparing them to a set of attribute values, the decision sample.

Decision by sampling is but one of several influential theories psychologists have embraced that suggest that comparisons play a central role in evaluating stimuli. Prospect theory—arguably the most influential of these—may have been influential in so many areas of the social sciences precisely because such comparisons play a limited role. That is, its comparison-based insights merely require identification of the neutral reference point, typically the status quo, around which a value function (a mapping of objective quantities to subjective value) is fit. In contrast, decision by sampling does not presume a preexisting value function but posits that all valuations arise from comparisons.

“We assume that the decision sample, to which a target...is compared, is a small, random sample...from memory” (Stewart et al., 2006, p. 4). The authors go on to say that “of course this random sampling assumption is likely to be incorrect” (p. 4). Subsequent research examined several properties of attribute values that make them more or less likely to enter decision samples. The present paper instead looked at how people *process* values to identify a possible mechanism by which values enter the decision sample. By answering this first-order mechanism question, the ambitious goals and full potential of DbS may be more fully realized.

We found that neither being exposed to nor having to respond to values is sufficient to place them in the decision sample. In other words, merely making values accessible—passively or through active engagement—does not consistently push them to be comparison standards that



guide subjective valuation<sup>4</sup>. Instead, subjectively *evaluating* values led them to enter that pool used to subjectively evaluate additional ones. Such patterns emerged in how people evaluated money (thus producing more or less loss aversion; Studies 1-2), time (thus producing more or less patience; Study 3), and vaccine efficacies (thus producing more or less vaccine hesitancy; Study 4). This cross-domain consistency is a promising sign that the present results offer a fairly general answer to the key open question of which attribute values serve as comparisons as people encounter and interpret attribute values.

Furthermore, Study 4 directly showed that evaluating (as opposed to merely responding to) values change the way that *other* values were subjectively evaluated (in a manner that DbS would anticipate), which explained the effects of the distribution of values participants evaluated on their interest in a novel vaccine. Especially given neither we (Supplemental Study B) nor previous researchers (Walasek & Stewart, 2019) have found that the memorability of values explains their inclusion in the decision sample, this data provided initial support for an alternative possibility. That is, it may not be that the decision sample is recruited at the time of decision-making to then guide such decisions through a comparison-based process. Instead, the process of subjectively evaluating values may then shape one's subjective sense of what values are relatively large or small, and thus attractive or unattractive, through the comparison-based ranking rules that DbS has shown to be crucial.

### **Implications, Future Directions, and Open Questions**

In what follows, we highlight next steps for a forward-looking research agenda, consider how the present findings have implications for those open questions, and further revisit—

---

<sup>4</sup> Studies 1 and 2 found only evaluation pushed values into the decision sample, but Study 3 showed that merely responding in light of an attribute value (by retyping it) had a smaller but statistically significant influence as well. Study 4 did not permit such a test.

considering the present findings—the very nature of the decision sample. In other words, we both draw attention to lingering uncertainties and consider how the present work helps to inform them:

**Duration.** Although evaluation inserts values into the decision sample, it remains unclear just how long they remain there. Instead of asking what leads numbers to enter into a decision sample, future research could ask what leads certain numbers to *exit* the decision sample. One possibility is that these values' membership in the sample simply fades with time. Another possibility is that the depth with which these values were evaluated—whether the values were subject to a cursory assessment or a more thorough analysis—may determine their longevity.

One hint as to which evaluated values will linger in the decision sample comes from previous research suggesting that extreme exemplars loom large in representations of the past (Gilbert & Wilson, 2007) and thus guide one's approach to and predictions about the future (Szpunar et al., 2018). For example, commuters asked to recall an instance in which they missed their train brought to mind an equally terrible memory as those asked to recall the *worst* instance in which they missed their train (Morewedge et al., 2005). Although those authors focused on this biased recall as a reason people may misforecast the extremity of the future, DbS identifies how this same phenomenon may influence evaluations of the present. Without many average values to serve as comparison standards, fluctuations in attribute values in the more middling range may seem especially unremarkable. For example, a marathoner who works to improve her average performance to the 75<sup>th</sup> percentile may feel this improvement was inconsequential if the finishing times that loom large in her decision sample (and thus define what is excellent versus terrible on her own subjective evaluation scale) are her best and worst finishes. Given the present paper's findings, extreme values may be precisely those that were particularly likely to be

subjectively evaluated (“Wow, I can’t believe how amazingly [fast / slow] I was!”), explaining why they emerge in the decision sample to guide thinking about the present and future. One practical implication is that encouraging people to subjectively evaluate more of their experiences may help them confront future decisions, opportunities, and outcomes in a more balanced, realistic way.

**Scope.** Another question relates to the breadth of previous evaluations from which the decision sample draws. Decision-makers are seemingly influenced by attribute values associated with similar targets (Bless & Schwarz, 2010; Rablen, 2008). But what constitutes a similar target? For example, although evaluating an airline ticket price from New York to Rome most obviously would involve comparisons with previous travels between those two cities, it would likely also include prices on other transatlantic routes (e.g., New York to Paris). One question is whether the reason why one is engaging in an evaluation in the first place affects whether the evaluated value lingers in the decision sample. For example, do the prices of Parisian hotels one has stayed in during business travel still enter the decision sample when one considers hotels for an upcoming holiday in the French capital? More generally, targets that merely reside in the same overall category (e.g., travel expenses) might be recruited into the decision sample. If so, might checked-luggage fees seem relatively inexpensive if high-priced airfare and hotel rates are relevant (travel-expense) comparison standards?

Some existing data suggests that the scope of values that inform the decision sample may be quite wide. An early demonstration of DbS leaned on the ranks of credit and debit amounts in bank accounts to explain asymmetries in how people responded to monetary gains and losses, respectively (Stewart et al., 2006). To take this evidence at face value, this suggests a potentially wide net that is cast in recruiting decision samples, one that draws on monetary transactions of

all types when evaluating monetary prospects. And other demonstrations have found that the recent salience of seemingly incidental values can influence decisions in essentially unrelated domains (Ungemach et al., 2011; cf. Matthews, 2012). But certainly, we could think of examples that stretch credulity: It seems people are unlikely to use a car's weight when subjectively characterizing a newborn's.

On this question, the present research's identification of evaluation as a core first-order mechanism helps to inform speculation. When people form subjective evaluations, they tend not to do so in a completely decontextualized manner. Instead, such evaluations occur against an implicit (or sometimes explicit) backdrop or frame of reference. People may say "That is an amazing price for a ticket to Europe" or "My dad had one of the slowest marathon times I've ever seen," suggesting that one is using a relatively more constrained (airfare for flights to Europe) or general (all marathon times) decision sample, respectively. If subjective evaluation is core to creating the decision sample, then a better understanding of the natural reference classes against which such evaluations occur may help to predict the scope of the decision sample.

**Spontaneous evaluation.** To know how to apply the present theoretical development to new contexts, more research must be done to understand which attribute values are *spontaneously* evaluated. That is, our studies used tightly controlled contexts to allow us to isolate the importance of evaluation to placing values in decision samples. But in naturalistic contexts, people are of course not confronted with experimental manipulations that ask them to subjectively characterize values. If people are making a decision in light of the perceived magnitude of a value, then such an evaluation should occur (and the similarity in results between our decision and evaluation conditions implies that it does), but when else?

Emotion researchers have long viewed affective reactions as quite basic (Zajonc, 1980), automatic (Ferguson & Zayas, 2009), and not reliant on higher-order cognitive mechanism (Hamm et al., 2003). More recently, Schneid et al. (2015) showed that much as people engage in spontaneous trait inferences (STIs) about others' personalities (Carlston et al., 1995; McCarthy & Skowronski, 2011), they form spontaneous evaluative inferences (SEIs) as well. Crucially, such inferences emerged to a similar extent regardless of whether the experimenter explicitly guided participants to form such impressions. In contrast, the very fact that we observed differences between our evaluation (when participants were explicitly asked to evaluate an attribute value) and, for example, response conditions (when participants were merely asked to retype the attribute value) suggests that such spontaneous evaluations are not inevitable. It thus seems that certain targets (e.g., social ones) invite more spontaneous evaluation than others.

Consider the finding that cultures differ in their reactions to death depending on the distribution of death tolls to which they are exposed by the media (Olivola & Sagara, 2009). It certainly does seem intuitive—and if the present paper is correct, it should be the case—that quantifiable tragedies, perhaps in part out of empathy, are events whose scope people spontaneously evaluate. Characterizations of mass fatalities as “unprecedented in number” or more limited tragedies as those that “certainly could have been worse” reflect the sort of subjective assessments that should place those values into decision samples. To continue with this example, when might spontaneous evaluation *not* occur? Although local media are disproportionately likely to consider local events, people also learn of global tragedies. One possibility is that local tragedies—given they may stir more interest and concern—are more likely to be subjectively evaluated than global ones. More generally, the present findings suggest

that understanding when and for what targets spontaneous evaluation occurs will help to determine when such comparisons will influence everyday judgment and decision-making.

**The nature of the decision sample.** Whereas all four studies focused on one mechanistic question (What places values in the decision sample?), Study 4 offered more direct evidence that evaluating values influences decision making because such evaluations change how other values are subjectively evaluated. This relates to the more fundamental question of the form in which the decision sample is represented. The answer to this question might have seemed intuitive: memory. By one understanding, the decision sample should include those values that emerge into or linger in working memory that then serve as comparison standards by which to interpret newly encountered values. And through this lens, the key question might seem to be whether evaluation requires sufficient depth of processing to make evaluated attribute values memorable and thus present in the decision sample ( Craik, 1973; Craik & Tulving, 1975).

But this seemingly straightforward line of argumentation is not compatible with two relevant studies. First, Walasek and Stewart (2019) found that variability in participants' memory for the (different distributions of) values that defined lotteries they encountered did not predict variation in how much participants showed DbS-consistent patterns of responses. Second, our own Supplemental Study B conceptually replicated this non-significant correlation and, furthermore, found that evaluating (vs. merely responding to) values did not increase memory for them. Of course, one possibility is that explicit instructions to recall values—as requests to scour one's memory stores—may not be a valid probe of which values are naturally accessible, unprompted, in memory and thus part of the decision sample.

That said, the lack of support for a memory-based interpretation suggests that potential comparison values (i.e., the decision sample) may not be recruited each time that a decision

needs to be made. After all, such repeated recruitment would be onerous. Instead, the process of evaluating new values may refine one's own updatable subjective numeric evaluation scale that describes what makes a relatively large or small (or relatively attractive or unattractive) value. This may be the very process that Study 4 probed more directly.

Study 4 did directly document that evaluating (as opposed to merely retyping) certain vaccine-efficacy values influenced one's vaccine hesitancy regarding vaccines characterized by *other* vaccine-efficacy values because those evaluations changed how these other vaccine efficacy rates were evaluated. Note that this understanding is not inconsistent with DbS and the comparison-based processes it emphasizes, but it could suggest that such comparisons may happen at an earlier stage (as newly evaluated values help to refine one's own subjective valuation scale) instead of at the moment of decision-making itself. It also explains why evaluation is the process that places values in the decision sample (given the decision sample itself is essentially the set of values used to define what is a relatively small, medium, or large value). This understanding may also explain why Wedell et al. (1987) found that faces served as comparison standards when they were presented (and evaluated) prior to, but not concurrently with, a target face. For such evaluations to encourage the updating of an internal scale that aids in the interpretation of newly encountered stimuli, then the comparison standard would need to instigate this process before the target stimulus is encountered and interpreted.

Finally, it would explain why values do not need to be especially memorable for them to influence the decision sample. Instead, it is only necessary that the vestiges of evaluating such values—and the accompanying effects on one's subjective evaluation scale itself—need to stay with the self. The recent college graduate who carefully considers and evaluates job offers with salaries of 45, 46, 48, 50, 53, and 65 thousand dollars develops a non-linear scale by which to

subjectively characterize attainable jobs as relatively low-paying (high 40s) and relatively high-paying (50s and 60s), even as the specific dollar amounts that informed that scale fade from memory. This interpretation would explain how encoding processes may shape the decision sample even though they do not influence attribute value retrieval.

**Norm theory.** Although decision by sampling is one prominent theory that aims to explain how attribute values are subjectively interpreted in light of other values, it is of course not the only such account. Given this, might the present findings inform adjacent theories that see a key role for comparisons in judgment and experience? Consider norm theory. It argues that when people encounter an object or event, such a probe calls to mind an evoked set whose elements are defined by multiple attributes that compose an availability profile that helps to define what is normal. Comparisons with that norm influence interpretation and experience of the present (Kahneman & Miller, 1986).

This suggests that instead of asking which previously encountered attribute values enter the decision sample, one could ask which previously encountered elements enter the availability profile to help define a counterfactual norm. Although norm theory provides its own parallel vocabulary to pose this question, features of norm theory suggest the answer may be more complicated. Whereas decision by sampling is ideally applied prospectively (In light of a decision sample, how will a newly encountered attribute value rank?), norm theory posits that norms are constructed *post hoc*. That is, the mutability of different features of an occurrence (e.g., causes more than effects, exceptional more than routine aspects, focal more than background elements, actions more than inactions) inform the backward reasoning process by which a stimulus elicits a norm. That said, much as the present work found that subjectively evaluated attribute values were more likely to serve as comparison standards in subjectively



evaluating new attribute values, future research may find that the extent to which previously encountered exemplars inform norms depends on a similar congruence between how those exemplars were processed at encoding and for what reason a norm is conjured.

**Cultural generality.** In considering the generality of our results, we have focused on the consistency of our findings across different judgment and decision-making contexts. But especially given our participants were all located in the United States, there is a parallel question of whether the psychology we captured can be exported to other cultural contexts. Given those who hail from more individualistic cultures like the U.S. have been shown to process stimuli in more segregated and less holistic or relational ways (e.g., Varnum et al., 2010), it may seem that those who hail from more interdependent cultures would display even more sensitivity to how focal attributes compare or relate to values in their decision samples. That said, given we established the key role that evaluation plays in producing such effects, a particularly relevant question is whether there are cultural differences in the degree to which people spontaneously evaluate stimuli. Although previous work has documented that Westerners are more likely to engage in spontaneous trait inferences than Easterners (Na & Kitayama, 2011; Shimizu et al., 2017), the different patterns of attention (Shimizu & Uleman, 2021) and the different attributional styles (Miyamoto & Kitayama, 2002) that mediate such effects presumably would not have implications for cultural variation in spontaneous evaluations. Regardless, much as previous research has noted that prospect theoretic parameters need to be tweaked when applied to new cultural contexts (e.g., Rieger et al., 2017; Wang et al., 2017), future research may find some culture-bound properties of how decision samples are built and applied.

## **Conclusion**

Psychologists have long appreciated that people make sense of the world based on their context. Decision by sampling formalizes how people both interpret attribute values and thus are influenced by them in forming judgments and decisions. Understanding which context cues guide this process requires moving beyond determining to which values people are merely exposed to instead learn which values people naturally evaluate. The long-term success of a domain-general theory of how comparisons guide subjective valuation requires a solid understanding of which attribute values serve as such comparisons.

### References

- Alempaki, D., Canic, E., Mullett, T. L., Skylark, W. J., Starmer, C., Stewart, N., & Tufano, F. (2019). Reexamining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation. *Management Science*, 65(10), 4841-4862.
- Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception & Psychophysics*, 49(4), 303-314.
- André, Q., & Langhe, B. De. (2021). No evidence for loss aversion disappearance and reversal in Walasek and Stewart (2015), *Journal of Experimental Psychology: General*, 12, 2659–2665.
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, 118(1), 73-106.
- Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173-96.
- Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1), 175–192.
- Bless, H., & Schwarz, N. (2010). Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. In *Advances in experimental social psychology* (Vol. 42, pp. 319-373). Academic Press.
- Boyce, C. J., Brown, G. D., & Moore, S. C. (2010). Money and happiness: Rank of income, not income, affects life satisfaction. *Psychological Science*, 21(4), 471-475.
- Brown, D. R. (1953). Stimulus similarity and the anchoring of subjective scales. *American Journal of Psychology*, 66, 199–214.

- Brown, G. D. A., & Matthews, W. J. (2011). Decision by sampling and memory distinctiveness: Range effects from rank-based models of judgment and choice. *Frontiers in Psychology*, 2(11), 1–4.
- Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3), 539–576.
- Brown, G. D. A., & Stewart, N. (2005). *Similarity sampling in judgment: Evidence against Range Frequency Theory*. Unpublished manuscript.
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3), 432–459.
- Callan, M. J., Shead, N. W., & Olson, J. M. (2011). Personal relative deprivation, delay discounting, and gambling. *Journal of Personality and Social Psychology*, 101(5), 955–973.
- Carlston, D. E., Skowronski, J. J., & Sparks, C. (1995). Savings in relearning: II. On the formation of behavior-based trait associations and inferences. *Journal of Personality and Social Psychology*, 69, 420–436. <https://doi.org/10.1037/0022-3514.69.3.429>.
- Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3), 268–294.
- Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 12(6), 599–607.
- Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction?. *Psychological Science*, 11(4), 280–284.

- De Dreu, C. K., & McCusker, C. (1997). Gain–loss frames and cooperation in two-person social dilemmas: A transformational analysis. *Journal of Personality and Social Psychology*, 72(5), 1093-1106.
- De Meza, D., & Dawson, C. (2021). Neither an optimist nor a pessimist be: Mistaken expectations lower well-being. *Personality and Social Psychology Bulletin*, 47(4), 540-550.
- DeVoe, S. E., House, J., & Zhong, C. B. (2013). Fast food and financial impatience: A socioecological approach. *Journal of Personality and Social Psychology*, 105(3), 476-494.
- Detweiler-Bedell, B., & Detweiler-Bedell, J. B. (2016). Emerging trends in health communication: The powerful role of subjectivism in moderating the effectiveness of persuasive health appeals. *Social and Personality Psychology Compass*, 10(9), 484-502.
- Duffy, J. F. (2004). Rethinking the prospect theory of patents. *University of Chicago Law Review*, 71, 439-510.
- Duffy, S., & Kitayama, S. (2007). Mnemonic context effect in two cultures: Attention to memory representations? *Cognitive Science*, 31, 1009-1020.
- Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, 51(4), 380-417.
- Eibach, R. P., & Keegan, T. (2006). Free at last? Social dominance, loss aversion, and White and Black Americans' differing assessments of racial progress. *Journal of Personality and Social Psychology*, 90(3), 453.
- Ersner-Hershfield, H., Garton, M. T., Ballard, K., Samanez-Larkin, G. R., & Knutson, B. (2009). Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving. *Judgment and Decision Making*, 4, 280-286.

- Ferguson, M. J., & Zayas, V. (2009). Automatic evaluation. *Current Directions in Psychological Science*, 18, 362-366.
- Ferh-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2020). Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, 40(2), 147–180.
- Frederick, S. (2003). Time preference and personal identity. In G. Loewenstein, D. Read, & R. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 89-113). New York: Russell Sage Foundation.
- Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.
- Gilbert, D. T., & Wilson, T. D. (2007). Propection: Experiencing the future. *Science*, 317, 1351-1354.
- Gill, D., Kissová, Z., Lee, J., & Prowse, V. (2019). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*, 65(2), 494-507.
- Greenwald, A. G., Abrams, R. L., Naccache, L., & Dehaene, S. (2003). Long-term semantic memory versus contextual memory in unconscious number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 29(2), 235.
- Guassi Moreira, J. F., McLaughlin, K. A., & Silvers, J. A. (2021). Characterizing the network architecture of emotion regulation neurodevelopment. *Cerebral Cortex*, 31(9), 4140-4150.
- Guthrie, C. (2003). Prospect theory, risk preference, and the law. *Northwestern University Law Review*, 97, 1115.

Hamm, A. O., Weike, A. I., Schupp, H. T., Treig, T., Dressel, A., & Kessler, C. (2003).

Affective blindsight: Intact fear conditioning to a visual cue in a cortically blind patient.

*Brain*, 126(2), 267-275.

Hershfield, H. E., & Bartels, D. M. (2018). The future self. In G. Oettingen, A. T. Sevincer, & P.

M. Gollwitzer (Eds.), *The psychology of thinking about the future* (pp. 89-109). The Guilford Press.

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T.

Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). The Guilford Press.

Holt, Charles A., Laury, S. K. (2005). Risk aversion and incentive effects : New data without order effects. *American Economic Review*, 95(3), 902–904.

Hornsey, M. J., Finlayson, M., Chatwood, G., & Begeny, C. T. (2020). Donald Trump and

vaccination: The effect of political identity, conspiracist ideation and presidential tweets on vaccine hesitancy. *Journal of Experimental Social Psychology*, 88, 103947.

Hornsey, M. J., Chapman, C. M., Alvarez, B., Bentley, S., Salvador Casara, B. G., Crimston, C.

R., ... & Jetten, J. (2021). To what extent are conspiracy theorists concerned for self versus others? A COVID-19 test case. *European Journal of Social Psychology*, 51(2), 285–293.

Hounkpatin, H. O., Wood, A. M., & Dunn, G. (2016). Does income relate to health due to

psychosocial or material factors? Consistent support for the psychosocial hypothesis requires operationalization with income rank not the Yitzhaki Index. *Social Science & Medicine*, 150, 76-84.

- Huttenlocher, J., Hedeges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 129(2), 220-241.
- Jolls, C., & Sunstein, C. R. (2006). Debiasing through law. *Journal of Legal Studies*, 35(1), 199-242.
- Jung, M. H., Critcher, C., & Nelson, L. D. (2023, October 25). Evaluations Are Inherently Comparative, But Are Compared To What? Retrieved from [osf.io/a9362](https://osf.io/a9362)
- Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12), 1625-1633.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5, 193–206.
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2), 136-153.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 363-391.
- Korobkin, R. (2012). Daniel Kahneman's Influence on Legal Theory. *Loyola University Chicago Law Journal*, 44(5), 1349–1356.
- Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 75(1), 23-55.
- Kunde, W., Kiesel, A., & Hoffmann, J. (2003). Conscious control over the content of unconscious cognition. *Cognition*, 88(2), 223-242.
- Laming, D. (1984). The relativity of "absolute" judgements. *British Journal of Mathematical and Statistical Psychology*, 37(2), 152-183.
- Laming, D. R. J. (1997). *The measurement of sensation*. Oxford University Press.



- Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76(2), 149-188.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6), 431.
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4), 1494-1502.
- Matthews, W. J. (2012). How much do incidental values affect the judgment of time? *Psychological Science*, 23(11), 1432-1434.
- McCaffery, E. J., Kahneman, D. J., & Spitzer, M. L. (1995). Framing the jury: Cognitive perspectives on pain and suffering awards. *Virginia Law Review*, 1341-1420.
- McCarthy, R. J., & Skowronski, J. J. (2011). You're getting warmer: Level of construal affects the impact of central traits on impression formation. *Journal of Experimental Social Psychology*, 47(6), 1304-1307.
- McDermott, R. (1998). Adolescent HIV prevention and intervention: A prospect theory analysis. *Psychology, Health & Medicine*, 3(4), 371-385.
- Mellers, B. A., & Cooke, A. D. J. (1994). Trade-offs depend on attribute range. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5), 1055-1067.
- Melrose, K. L., Brown, G. D., & Wood, A. M. (2013). Am I abnormal? Relative rank and social norm effects in judgments of anxiety and depression symptom severity. *Journal of Behavioral Decision Making*, 26(2), 174-184.
- Mercer, J. (2005). Prospect theory and political science. *Annual Review of Political Science*, 8, 1-21.

- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity. *Journal of Personality and Social Psychology*, 83, 1239–1248.
- Morewedge, C. K., Gilbert, D. T., & Wilson, T. D. (2005). The least likely of times: How remembering the past biases forecasts of the future. *Psychological Science*, 16(8), 626-630.
- Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes*, 98(2), 133-143.
- Na, J., & Kitayama, S. (2011). Spontaneous trait inference is culture-specific: Behavioral and neural evidence. *Psychological Science*, 22, 1025-1032.
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. Freeman.
- Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review*, 125(4), 512–544.
- Olivola, C. Y., & Sagara, N. (2009). Distributions of observed death tolls govern sensitivity to human fatalities. *Proceedings of the National Academy of Sciences*, 106(52), 22151–22156.
- Oyserman, D., & Lee, S. W. S. (2008). A situated cognition perspective on culture: Effects of priming cultural syndromes on cognition and motivation. In R. Sorrentino & S. Yamaguchi (Eds.), *Handbook of motivation and cognition across cultures* (pp. 237–265). New York, NY: Elsevier.

- Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs*, 77, (92, Whole No. 565).
- Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*, 219(6), 84-93.
- Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler, H. F. J. M. Buffort, E. L. J. Leeuwenberg, & V. Sarris (Eds.), *Modern issues in perception* (pp. 262-282). VEB Deutscher Verlag der Wissenschaften.
- Qian, J., & Brown, G. D. A. (2005). Similarity-Based Sampling : Testing a Model of Price Psychophysics Development of Price Psychophysics. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 27(27), 1785–1790.
- Rablen, M. D. (2008) Rablen, M. D. (2008). Relativity, rank and the utility of income. *The Economic Journal*, 118(528), 801-821.
- Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, 253(5023), 980-986.
- Real, L. A. (1996). Paradox, performance, and the architecture of decision-making in animals. *American Zoologist*, 36(4), 518-529.
- Rieger, M. O., Wang, M., & Hens, T. (2017). Estimating cumulative prospect theory parameters from an international survey. *Theory and Decision*, 82(4), 567-596.
- Rossen, I., Hurlstone, M. J., & Lawrence, C. (2016). Going with the grain of cognition: applying insights from psychology to build support for childhood vaccination. *Frontiers in Psychology*, 7, 1483.
- Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: the role of message framing. *Psychological Bulletin*, 121(1), 3-19.

- Schneid, E. D., Carlston, D. E., & Skowronski, J. J. (2015). Spontaneous evaluative inferences and their relationship to spontaneous trait inferences. *Journal of Personality and Social Psychology, 108*(5), 681-696.
- Schwarz, N., Münnkel, T., & Hippler, H. J. (1990). What determines a ‘‘perspective’’? Contrast effects as a function of the dimension tapped by preceding question. *European Journal of Social Psychology, 20*, 357–361.
- Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology, 55*(2), 150-155.
- Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: a commentary on capacity limitations. *Psychological Review, 101*(2), 357-361.
- Shimizu, Y., Lee, H., & Uleman, J. S. (2017). Culture as automatic processes for making meaning: Spontaneous trait inferences. *Journal of Experimental Social Psychology, 69*, 79-85.
- Shimizu, Y., & Uleman, J. S. (2021). Attention allocation is a possible mediator of cultural variations in spontaneous trait and situation inferences: Eye-tracking evidence. *Journal of Experimental Social Psychology, 94*, 104115.
- Skrynka, J., & Vincent, B. T. (2019). Hunger increases delay discounting of food and non-food rewards. *Psychonomic Bulletin & Review, 26*(5), 1729-1737.
- Steinacker, A. (2006). Externalities, prospect theory, and social construction: When will government act, what will government do? *Social Science Quarterly, 87*(3), 459-476.
- Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive psychology, 53*(1), 1–26.
- Szpunar, K. K., Shrikanth, S., & Schacter, D. L (2018). Varieties of future-thinking. In G.

- Oettingen, A. T. Sevincer, & P. M. Gollwitzer (Eds.), *The psychology of thinking about the future* (pp. 52-67). Guilford Publications.
- Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy*, 112(S1), S164-S187.
- Thorndyke, P. W. (1981). Distance estimation from cognitive maps. *Cognitive Psychology*, 13, 526-550.
- Tripp, J., & Brown, G. D. A. (2016). Being paid relatively well most of the time: Negatively skewed payments are more satisfying. *Memory & Cognition*, 44(6), 966–973.
- Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. In G. Wright (Eds), *Behavioral decision making* (pp. 25-41). Springer.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297-323.
- Tversky, A., & Koehler, D. J. (1994). Support theory: a non-extensional representation of subjective probability. *Psychological Review*, 101(4), 547-567.
- Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science*, 22(2), 253-260.
- Varnum, M. E. W, Grossmann, I., Kitayama, S., & Nisbett, R. E. (2010). The Origin of cultural differences in cognition: Evidence for the social orientation hypothesis. *Current Directions in Psychological Science*, 19(1), 9-13.
- Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, 15(11), 546-554.

- Volkman, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 273-294). Harper & Row.
- Volkman, J. (1951). Scales of judgment and their implications for social psychology. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads* (pp. 273–294). New York, NY: Harper.
- Walasek, L., & Brown, G. D. (2015). Income inequality and status seeking: Searching for positional goods in unequal US states. *Psychological Science*, 26(4), 527-533.
- Walasek, L., Mullett, T. L., & Stewart, N. (2021). Acceptance of mixed gambles is sensitive to the range of gains and losses experienced, and estimates of lambda ( $\lambda$ ) are not a reliable measure of loss aversion: Reply to André and de Langhe (2021). *Journal of Experimental Psychology: General*, 150(12), 2666-2670.
- Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, 144(1), 7–11.
- Walasek, L., & Stewart, N. (2019). Context-dependent sensitivity to losses : Range and skew manipulations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(6), 957–968.
- Walasek, L., & Stewart, N. (2021). You cannot estimate an individual's loss aversion using an accept-reject task, *Decision*, 8, 2-15.
- Walker, J., Risen, J. L., Gilovich, T., & Thaler, R. (2018). Sudden-death aversion: Avoiding superior options because they feel riskier. *Journal of Personality and Social Psychology*, 115(3), 363-378.

- Wang, M., Rieger, M. O., & Hens, T. (2017). The impact of culture on loss aversion. *Journal of Behavioral Decision Making*, 30(2), 270-281.
- Wedell, D. H., Parducci, A., & Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, 23, 230-249.
- Wood, A. M., Brown, G. D., Maltby, J., & Watkinson, P. (2012). How are personality judgments made? A cognitive model of reference group effects, personality scale responses, and behavioral reactions. *Journal of Personality*, 80(5), 1275-1311.
- Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2), 151-175.

## **SUPPLEMENTAL MATERIALS**

### **Evaluations Are Inherently Comparative, But Are Compared To What?**

Minah H. Jung<sup>1</sup>, Clayton R. Critcher<sup>2</sup>, Leif D. Nelson<sup>2</sup>

<sup>1</sup>New York University, <sup>2</sup>University of California, Berkeley

Page 2: Study A

Page 6: Study B

Page 10: Additional analyses of all data using different exclusion criteria.



## Study A

Study A tested whether exposure was sufficient to place values into the decision sample. All participants saw lotteries defined over the same narrow range of losses (-\$20 to -\$6). What we varied was the range of *gains* to which participants were exposed (narrow or wide) as well as whether participants made decisions over the narrow or wide range of gain values. Those in the *wide* gain range condition made decisions about lotteries whose gains ranged from \$6 to \$32. Those in the *narrow* gain range condition made decisions about lotteries whose gains ranged from \$6 to \$20.

DbS is clear that the loss aversion coefficient should be higher in the wide condition compared to narrow condition. But is this merely (or partly) because those in the wide condition were *exposed* to a wider range of gain values? A third condition was instrumental in answering that question. In a *narrow + exposure* condition, participants made decisions about lotteries over the narrow gain range (\$6 to \$20) but were exposed to the full range of gain values (\$6 to \$32). If exposure is sufficient to place values into the decision sample, then the narrow + exposure condition should prompt a higher loss aversion coefficient than the narrow condition. If instead exposure is not sufficient to insert values into the decision sample, the loss aversion coefficient for those in the narrow + exposure condition should be similar to those in narrow condition (and thus smaller than those in the wide condition).

## Method

**Participants and design.** Participants were 1,070 Americans recruited through Amazon Mechanical Turk (AMT). They were randomly assigned to one of three gain range conditions: wide, narrow, or narrow + exposure.

**Procedure.** All participants saw lotteries defined by the same narrow range of losses (-\$20 to -\$6, in \$2 increments: -\$20, -\$18, ..., -\$8, -\$6). We modified the way that lotteries were presented in Walasek and Stewart (2015) to decouple the range of values used to define lotteries from the range of values to which participants were exposed. On every trial, participants were exposed to the same two number lines—one for gains, one for losses. The gain and loss values for a particular lottery were identified on their respective number line (see Figure S1).

We varied the range of values to which participants were exposed by varying the width of the number line. The gain number line spanned from \$6 to \$20 in the narrow condition (Figure S1A), but from \$6 to \$32 in the wide condition (Figure S1B). Although the gain values of the lotteries in the narrow + exposure condition ranged only from \$6 to \$20 (as in the narrow condition), the number line ranged from \$6 to \$32 (as in the wide condition). All participants saw a loss number line that ranged from -\$20 to -\$6.

Those in the wide gain range condition indicated whether they would accept or pass on the 112 unique lotteries that could be defined by every combination of the 14 gain and 8 loss values. In contrast, those in the narrow and narrow + exposure conditions saw the 64 unique lotteries that could be created by every combination of the 8 gain and 8 loss values. In order to roughly equate across the number of lotteries that participants saw, these participants responded to these 64 lotteries twice. The order of the lotteries was randomized.

## **Results and Discussion**

To determine whether our manipulations influenced participants' degree of loss aversion, we first had to calculate participants' loss aversion coefficients. For each participant, we conducted a logistic regression in which we predicted a particular participant's decision to accept (+1) or reject (-1) a lottery as a function of the gain value and the loss value of the lottery. From

**Figure S1**

*An Example Lottery as Seen by Those in the Narrow Condition (Panel A) and Wide and Narrow + Exposure Conditions (Panel B) in Study A*

(A)



(B)



this regression, we took the beta for the loss value and divided it by the beta for the gain value. This quotient, once multiplied by negative one, reflects participants' relative sensitivity to losses vs. gains when considering risky decisions (see Walasek & Stewart, 2015).

As in Studies 1 and 2, we precisely followed Walasek and Stewart's (2015) exclusion criteria, which we describe next. First, we excluded participants with incomplete responses. Second, we excluded those participants whose regression fit deviance scores were among the remaining highest 5%. Third, we excluded those who displayed a negative loss aversion coefficient. Such participants indicated greater interest in lotteries with lower expected values, thereby indicating a failure to understand the procedure or take it seriously. This left 868 participants in all analyses reported below. Conceptually replicating the findings in Walasek and Stewart's (2015), we found that participants exposed to a wide range of gains showed greater loss aversion (*Median coefficient* = 1.13) compared to those who saw a narrow range of gains (*Median coefficient* = 1.03),  $Z = 3.35$ ,  $p < .001$ .

Was exposure to a wide range of gains sufficient to place those values in the decision sample, thereby inflating the loss aversion coefficient? In a word, no. Those in the narrow + exposure condition showed a relatively low loss aversion coefficient (*Median* = 1.02), roughly comparable in size to that of the narrow condition,  $Z = 0.95$ ,  $p = .343$ . This loss aversion coefficient was significantly smaller than displayed by those who were not merely exposed to but

made lottery decisions over the wide range of gains,  $Z = 4.16$ ,  $p < .001$ . Given this lack of support for the exposure account, we use narrow + exposure as a baseline comparison condition in both Studies 1 and 2.

### Study B

In Studies 1 and 2, the trials that delivered the key manipulation—i.e., those lotteries including gain values above \$20 that participants were merely exposed to, responded in light of, evaluated, or made decisions about—were interspersed with those trials whose decisions the manipulation might shape—i.e., those lotteries with gain values of \$20 or less. This means that the manipulation could not influence responses on the first few trials. This feature also characterized the work of Walasek and Stewart (2015), the motivation behind Studies 1 and 2's paradigms. Furthermore, the de facto delayed introduction of the manipulation should have only made our focal tests more conservative.

Studies 3 and 4 avoided this timing-related limitation by having all participants experience the key manipulation before responding to the focal trials. Study 3 tested our ideas in the context of patience, and Study 4 tested our ideas in the context of vaccine hesitancy. In Study B, we introduce the same procedural change in an investigation of loss aversion. Much like in Studies 3 and 4, participants in Study B were first exposed to trials introducing the key manipulation. That is, all participants first encountered gain values from \$22 to \$32. Participants in the *response* condition retyped those values, whereas participants in the *evaluation* condition subjectively rated the value's attractiveness. Participants all went on to make decisions over the same set of gain values (i.e., those from \$6 to \$20). Those decisions were used to calculate loss aversion coefficients for each participant. If evaluation places values in the decision sample (as

Studies 1-4 all suggested), we should expect to see a larger loss aversion coefficient among those in the evaluation (compared to the response) condition.

Study B also included a second novel feature. At the study's conclusion, all participants were asked to recall as many gain values as they could. Walasek and Stewart (2019) found that those who had a better memory for the values that defined lotteries did not show any more or less loss aversion (even though manipulation of those values did influence loss aversion, as indexed by AUIC), suggesting that such memory measures do not recover the contents of the decision sample. Regardless, our exploratory inclusion of this measure permitted us to independently test whether such recall differed by condition, as well as whether recall rates predicted the influence of the manipulation. This would offer an independent test of whether such memory measures are useful in uncovering the decision sample.

## **Method**

**Participants and design.** We recruited 911 Americans from AMT. Participants were randomly assigned to one of two gain range conditions: response or evaluation. Following the same exclusion criteria specified by Walasek and Stewart (2015) and used in our earlier loss aversion studies, we included the remaining 783 participants in our analyses.

**Procedure.** The procedure was similar to Studies 1 and 2. The response condition was nearly identical to the narrow + response in Studies 1 and 2: Participants made decisions about whether to accept or reject lotteries whose gain values ranged from \$6 and \$20. For lotteries with gain values between \$22 and \$32, they retyped those gain values. The evaluation condition was nearly identical to the narrow + gain evaluation condition in Study 2: Participants in the evaluation condition made accept or reject decisions over the lotteries with gain values between

\$6 and \$20, but for the other lotteries, they rated the attractiveness of the gain values between \$22 and \$32 on a slider scale that ranged from “not at all attractive” to “extremely attractive.”

To ensure that responding to (i.e., retyping) or evaluating gain values had a chance to modify participants’ decision samples before their very first lottery trials—analogous to what was done in Studies 3 and 4—we had all participants first consider the lotteries with the high gain values between \$22 and \$32 (i.e., those specific to each condition and that would be used on the subsequent response or evaluation trials) before beginning the main task. Response participants had to retype the 6 gain values. Evaluation participants had to indicate how attractive it would be to win each of the 6 gain values. These participants responded on non-numeric slider scales that ranged from “not at all attractive” to “extremely attractive.” At that point, the 112 lotteries appeared in random order.

We then examined whether our manipulation influenced participants’ recall of gain values they considered, which might mediate the influence of our manipulations of loss aversion. After completing their lottery task, participants were asked to “try to recall all of the different amounts of money that were presented as possible win values.” Participants saw 20 blank boxes and were told to “type each unique win amount that you can remember in the boxes below. Please type one value per box... There may be more boxes below than there were winning values, so you should not feel like you need to put an answer in each one.”

## Results

**Loss aversion.** We calculated loss aversion coefficients and trimmed the data using the same logistic regression procedures described previously. One hundred twenty-eight participants were excluded based on Walasek and Stewart’s (2015) exclusion criteria. This left 783

participants for the final analyses. See additional analyses with different exclusion criteria in the Additional Analyses section below.

As predicted, we found that participants displayed greater loss aversion when they evaluated the attractiveness of high gain values (*Median coefficient* = 1.18, 95% bootstrapped CI = [1.12, 1.25]) compared to when they merely responded to those gain values (*Median coefficient* = 1.02, 95% bootstrapped CI = [1.00, 1.05]),  $Z = 3.80, p < .001$ . The AUIC approach showed similar findings. We saw greater evidence of loss-averse behavior in the evaluation (*Median AUIC* = 0.33) compared to the response condition (*Median AUIC* = 0.45),  $Z = 6.22, p < .001$ . This means that regardless of whether our manipulation had a chance to modify participants' decision sample before (the present study) or just after they started the lottery task (Studies 1-2), evaluating the high gain values produced greater loss aversion than merely responding by retyping them. This provides further evidence that evaluation places values in the decision sample.

**Memory.** For participants' recall of gain values, we coded each recalled gain value according to whether it was one of the 14 gain values that did indeed appear in the task or not. Even for those participants who survived Walasek and Stewart's (2015) exclusion criteria, eight of the participants dropped out without completing this gain-value recall measure. Participants in the evaluation condition did not recall any more gain values ( $M = 7.52, SD = 3.16$ ) than those in the response condition ( $M = 7.56, SD = 2.95$ ),  $t < 1$ . Furthermore, the number of gain values participants accurately recalled did not correlate with their loss aversion coefficient,  $r(773) = -0.02, p = .562$ . AUIC also showed no correlation with memory  $r(773) = -.06, p = .097$ . In other words, the evaluation manipulation neither enhanced memory for gain values, nor did memory for gain values significantly correlate with loss aversion.



## Additional Analyses

In this section, we report additional analyses, first using a laxer and then a more stringent inclusion criterion. For example, we report analyses including all participants from Studies A and B and 1-4. We also report analyses similar to those reported in the main text (or in the original reports of Supplemental Studies A and B) but with a sample further winnowed with the use of a memory test. That task restricted the included sample only to those who demonstrated accurate recall. Although certain effects reported as “trending” in the main text become significant (or vice versa), the output of these reanalyses essentially bolster the robustness of the conclusions drawn in the main manuscript.

### Study A: Analyses Including All Participants

Participants exposed to a wide range of gains showed greater loss aversion (*Median coefficient* = 1.12; 95% bootstrapped CI = [1.04, 1.22]) compared to those who saw a narrow range of gains (*Median coefficient* = 1.02, 95% bootstrapped CI = [1.00, 1.04]),  $Z = 3.60$ ,  $p < .001$ .

Those in the narrow + exposure condition showed a relatively low loss aversion coefficient (*Median coefficient* = 1.02, 95% bootstrapped CI = [1.00, 1.04]), roughly comparable in size to that of the narrow condition,  $Z = 0.69$ ,  $p = .489$ . This loss aversion coefficient was clearly smaller than the coefficient of those in the wide condition,  $Z = 4.24$ ,  $p < .001$ .

### Study B

**Analyses including all participants.** Participants displayed greater loss aversion when they evaluated the attractiveness of high gain values (*Median coefficient* = 1.07, 95% bootstrapped CI = [1.03, 1.16]) compared to when they merely responded to those gain values (*Median coefficient* = 1.00, 95% bootstrapped CI = [1.00, 1.03]),  $Z = 2.38$ ,  $p = .017$ . A total of

900 participants completed the gain-value recall measure. Participants in the evaluation condition did not recall more gain values ( $M = 7.23$ ,  $SD = 3.27$ ) than those in the response condition ( $M = 7.33$ ,  $SD = 3.12$ ),  $t < 1$ . Furthermore, the number of gain values participants accurately recalled only marginally correlated with their loss aversion coefficient,  $r(898) = .08$ ,  $p = .084$ .

**Memory question.** Although many researchers use attention check questions that determine whether participants are indeed paying attention *in the moment* (e.g., “Please respond 4 on this question”), we included a more difficult memory question at the study’s end. Such memory questions—because they require that participants were both paying attention earlier in the study *and* still remember that detail later—produce greater failure rates. That said, Jung, Fausto, and Critcher (2020) found that such questions—despite the higher failure rates they produce than standard attention checks—can be quite effective screening mechanisms for distinguishing those participants who are carefully engaged in a study versus not. Jung et al. (2020) show that steps taken to improve performance on such memory questions can simply reduce such questions’ ability to screen for inattentive participants without any associated benefit on data quality.

The memory question was, “You saw 112 lotteries. For roughly half of them, you indicated whether you would be willing to take the gamble by clicking Accept or Reject. What did you do for the other lotteries?” The accurate answer depended on the condition participants were randomly assigned to: I typed how much I would win if the coin landed on Heads, I evaluated how attractive the outcome would be if the coin landed on Heads, I explained how I would spend the money if I won, or I forecast how much regret I would experience if I lost.

**Analyses excluding those who failed to answer the memory question correctly or failed to meet the exclusion criteria by Walasek & Stewart (2015).** Of the 783 participants who

survived Walasek and Stewart (2015) exclusion criteria, 775 of them attempted the memory question that followed the gain-value recall task. One hundred thirty of these participants answered the moderately difficult memory question incorrectly. Below, we report analyses with the remaining 645 participants.

Participants displayed greater loss aversion when they evaluated the attractiveness of high gain values (*Median coefficient* = 1.19, 95% bootstrapped CI = [1.11, 1.25]) compared to when they merely responded to those gain values (*Median coefficient* = 1.01, 95% bootstrapped CI = [1.00, 1.04]),  $Z = 3.91, p < .001$ . A total of 645 participants correctly answered the memory question. Participants in the evaluation condition did not recall any more gain values ( $M = 7.58, SD = 3.07$ ) than those in the response condition ( $M = 7.81, SD = 2.90$ ),  $t < 1$ . Furthermore, the number of gain values participants accurately recalled did not correlate with their loss aversion coefficient,  $r(643) = .00, p = .974$ .

## Study 1

**AUIC analyses.** AUIC is tied to the number of lotteries participants choose to accept, with greater AUIC reflecting diminished loss aversion. We calculated AUIC using the same set of lotteries for all participants, those with gain values between \$6 and \$20. The results using AUIC replicated our just-reviewed analyses that used logistic regressions to calculate the loss aversion coefficient. More specifically, participants who made decisions over the wide range of gains displayed a similar AUIC (*Median* = .36; 19.89 of 64 lotteries accepted, on average) as those in the narrow + evaluation condition (*Median* = .39; 22.15 of 64 lotteries accepted, on average),  $Z = 1.00, p = .318$ . But those in the narrow + evaluation condition did have a lower AUIC than those in the narrow + response condition (*Median* = .45; 25.64 of 64 lotteries accepted, on average),  $Z = 4.21, p < .001$ . Those in the narrow + response condition did not have

an even high AUIC compared to those in the narrow + exposure condition (*Median* = .45; 26.48 of 64 lotteries accepted, on average),  $Z < 1$ . In other words, the AUIC analyses again suggest that evaluation is what places attribute values into the decision sample (see Figure 3).

**Analyses including all participants.** Participants in the wide condition showed more loss aversion (*Median coefficient* = 1.26, 95% bootstrapped CI = [1.15, 1.37]) than those in the narrow + exposure condition (*Median coefficient* = 1.00, 95% bootstrapped CI = [1.00, 1.03]),  $Z = 4.58, p < .001$ .

Participants in the narrow + response condition displayed significantly less loss aversion than those in the wide condition,  $Z = 5.73, p < .001$ . And they actually showed marginally less loss aversion (*Median coefficient* = 1.00, 95% bootstrapped CI = [1.00, 1.00]) than those in the narrow + exposure condition,  $Z = 1.70, p = .089$ . Although this difference is only marginal, note that it goes in the opposite direction from what would be expected if responding was itself sufficient to place values in the decision sample.

Participants in the narrow + evaluation condition had a higher loss aversion coefficient (*Median coefficient* = 1.08, 95% bootstrapped CI = [1.02, 1.13]) than those in the narrow + response condition,  $Z = 3.51, p = .001$ , as well as those in the narrow + exposure condition,  $Z = 2.08, p = .038$ . Though narrow + evaluation participants—at least when calculating the loss aversion coefficient over all lotteries—did display less loss aversion than those in the wide condition (*Median coefficient* = 1.26, 95% bootstrapped CI = [1.15, 1.37]),  $Z = 2.38, p = .017$ . But was that simply because the wide range condition included a different set of lotteries over which the loss aversion coefficient was calculated?

We next calculated a loss aversion coefficient for participants using only those decisions made by participants in all four conditions. More specifically, this reanalysis includes only those

lotteries with gain values ranging from \$6 to \$20, inclusive. And indeed, with analyses now calculated over an equivalent set of lotteries, narrow + evaluation participants displayed similar loss aversion to those participants in the wide condition (*Median coefficient* = 1.16, 95% bootstrapped CI = [1.03, 1.29]),  $Z = 1.21$ ,  $p = .227$ . This replicates the finding reported in the main manuscript that the wide condition only seems to elevate loss aversion compared to the narrow + evaluation condition because—in the original analyses—the wide condition’s loss aversion coefficient was calculated over more (and thus different) lotteries.

## Study 2

**AUIC analyses.** Participants who made decisions over the wide range of gains had a similar AUIC (*Median* = .34; 18.81 of 64 lotteries accepted, on average) as those in the narrow + lottery evaluation condition (*Median* = .34; 19.05 of 64 lotteries accepted, on average),  $Z < 1$ , and those in the narrow + gain evaluation condition (*Median* = .34; 18.33 of 64 lotteries accepted, on average),  $Z < 1$ . Furthermore, the two evaluation conditions were not statistically different,  $Z = 1.07$ ,  $p = .283$ . Those in the narrow + exposure condition had a higher AUIC (*Median* = .45; 25.81 of 64 lotteries accepted, on average) than those in the three other conditions,  $Zs > 8.30$ ,  $ps < .001$ .

**Analyses including all participants.** Participants were more loss averse in the wide condition (*Median coefficient* = 1.39, 95% bootstrapped CI = [1.26, 1.51]) than in the narrow + exposure condition (*Median coefficient* = 1.05, 95% bootstrapped CI = [1.03, 1.08]),  $Z = 7.69$ ,  $p < .001$ . Furthermore, participants who subjectively evaluated lotteries (i.e., the narrow + lottery evaluation condition) showed more loss aversion (*Median coefficient* = 1.22, 95% bootstrapped CI = [1.16, 1.28]) than those in the narrow + exposure condition,  $Z = 3.87$ ,  $p < .001$ . Narrow +

lottery evaluation participants showed less loss aversion than those in the wide condition,  $Z = 3.49, p < .001$ .

Participants in the narrow + gain evaluation condition were substantially more loss averse (*Median coefficients* = 1.23, 95% bootstrapped CI = [1.17, 1.32]) than those in the narrow + exposure condition,  $Z = 4.55, p < .001$ . But—like narrow + lottery evaluation participants—they were less loss averse than those in the wide condition,  $Z = 2.84, p = .005$ . Participants' loss aversion did not differ between the two evaluation conditions,  $Z = 0.67, p = .502$ .

Did participants show more loss aversion in the wide condition than in the two gain conditions because loss aversion was calculated over different lotteries, or because actually making decisions using values (i.e., gains between \$22 and \$32) more strongly placed them in the decision sample? In short, the former was the case. We concluded this upon reconducting analyses of the wide condition using only those lotteries that were used in the two evaluation conditions—i.e., those with gains ranging from \$6 to \$20, inclusive. Participants' loss aversion in the wide condition did not significantly differ from participants' loss aversion in the narrow + lottery evaluation condition,  $Z = 0.64, p = .521$ , or the narrow + gain evaluation condition,  $Z = 1.37, p = .170$ . In other words, it was only evaluation—of either the lottery or the gain—that placed values in the decision sample. The apparent additional influence of actually making decisions was instead merely an artifact of wide participants making decisions over a different set of lotteries.

**Memory question.** Study 2 also included a memory question asking participants, “Different participants are asked to do different things in this study. Which most accurately describes what you were asked to do?” Participants had to select one of four options. The accurate answer depended on the condition to which participants were randomly assigned:

- For all lotteries, you indicated whether you would accept or reject them.
- For some lotteries (but not all), you rated the attractiveness of the outcome if the lottery came up heads (instead of the lottery as a whole)
- For some lotteries (but not all), you rated the attractiveness of the lottery as a whole.
- For some lotteries (but not all), you rated the attractiveness of the outcome if the lottery came up tails (instead of the lottery as a whole).

Note that this question was difficult. It required that participants not merely remember specifics of the task that they completed (e.g., whether they rated the attractiveness of a lottery or an outcome), but also that they remember whether they were making judgments about whether Heads or Tails were to be flipped. Although such difficult questions may serve as effective screening mechanisms for who is fully engaged (Jung et al., 2020), we did not use the question in this way when presenting results in the main text (so that readers would not be left with the mistaken impression that our results depended on such exclusions).

**Analyses excluding those who failed to answer the memory question correctly or failed to meet the exclusion criteria by Walasek & Stewart (2015).** Two hundred seventeen participants were excluded based on Walasek and Stewart's (2015) exclusion criteria. An additional 549 participants failed to accurately answer the memory question. We conducted analyses on the remaining 1,457 participants.

We again found that participants displayed greater loss aversion when they made decisions over a wide range of gains (*Median coefficient* = 1.31, 95% bootstrapped CI = [1.20, 1.44]) compared to a narrow range of gains (*Median coefficient* = 1.08; 95% bootstrapped CI = [1.05, 1.13]),  $Z = 6.58$ ,  $p < .001$ . Furthermore, we replicated the findings from Study 1 that evaluating the attractiveness of the wider range of lotteries (narrow + lottery evaluation) also

elevated loss aversion (*Median coefficient* = 1.31; 95% bootstrapped CI = [1.24, 1.44]) compared to the narrow + exposure condition,  $Z = 4.46$ ,  $p < .001$ .

Participants in the new narrow + gain evaluation condition showed elevated loss aversion (*Median coefficient* = 1.25; 95% bootstrapped CI = [1.19, 1.35]) compared to the narrow + exposure condition,  $Z = 3.68$ ,  $p < .001$ . Furthermore, the two evaluation conditions were not statistically distinguishable,  $Z = 0.63$ ,  $p = .527$ .

Finally, we recalculated the loss aversion coefficients using only those lotteries that all participants accepted or rejected—i.e., those with gain values ranging from \$6 to \$20, inclusive. This reduced the loss aversion observed in the wide condition (*Median coefficient* = 1.30; 95% bootstrapped CI = [1.19, 1.44]) so that it was no longer greater than the loss aversion observed in the narrow + lottery evaluation,  $Z = 0.95$ ,  $p = .344$ , or the narrow + gain evaluation conditions,  $Z = 0.25$ ,  $p = .799$ .

### Study 3

**Memory question.** Study 3 also used a memory question to effectively screen for people who were sufficiently engaged (c.f., Jung et al., 2020). The question read: “In this study, you saw 60 pairs of payoff options. Different participants are asked to do a different task with these pairs of payoff options. Which most accurately describes what you were asked to do?” The accurate answer depended on the condition participants were randomly assigned to:



- For all pairs of payoff options, I made choices about which of one of the two options I prefer.
- For some pairs of payoff options, I made choices about which one of the two I prefer, but for the other pairs, I typed in the lengths of time I would have to wait to receive the larger payoff in a blank space
- For some pairs of payoff options, I made choices about which one of the two I prefer, but for the other pairs, I indicated how unappealing it would be to wait for a certain amount of time to receive the larger payoff.
- For some pairs of payoff options, I typed in the lengths of time I would have to wait to receive the larger reward in a blank space, but for the other pairs, I indicated how unappealing it would be to wait for a certain amount of time to receive the larger payoff.

As before, although results in the main text were reported including those who missed this memory question, we reconducted analyses on the smaller subset of participants who answered this memory question correctly.

**Analyses excluding those who failed to answer the memory question correctly.** We excluded 229 participants who were unable to answer the memory question correctly. All analyses were conducted on the remaining 980 participants.

We began by testing whether those exposed to the uniform distribution (uniform participants) of time delays displayed more patience than those who saw the skewed distribution (skewed participants). After all, the required delays for the larger reward should have seemed subjectively shorter when considered in the context of the uniform than the skewed distribution. And indeed, this first prediction was confirmed. When considering the same tradeoffs, uniform participants indicated a willingness to wait for the larger reward on more trials (55.30%) than did skewed participants (50.50%),  $t(976.98) = 2.41, p = .016$ .

Did such patience take the form that DbS would anticipate? More specifically, we expected that uniform participants' greater patience would emerge most clearly at 2 months

(when the rankings were highly discrepant between conditions) and least strongly at 12 months (when the rankings were barely discrepant between conditions). We coded time delay as -1 (2 months), 0 (6 months), and +1 (12 months). As expected, we observed a Time Distribution  $\times$  Time Delay interaction,  $t(28419) = 8.12, p < .001$ . This reflected the expected pattern. Uniform participants were much more patient at 2 months,  $t = 4.34, p < .001$ ; less so at 6 months,  $t = 2.41, p = .016$ ; and did not differ significantly from skewed participants at 12 months,  $t < 1$ .

To probe our central question—whether evaluation places values in the decision sample—we tested whether this finding was driven by participants in the evaluation (compared to the response) condition. And indeed, the Time Distribution  $\times$  Time Delay  $\times$  Task interaction was significant,  $t(28414.99) = 5.03, p < .001$ . When participants subjectively evaluated the additional values, a significant Time Distribution  $\times$  Time Delay interaction suggested that such values populated the decision sample and influenced patience as decision by sampling would predict,  $t(28414.99) = 10.01, p < .001$ . In contrast, when participants merely retyped the additional values, this Time Distribution  $\times$  Time Delay interaction was much weaker,  $t(28414.99) = 2.51, p = .012$ .

## Study 4

**Memory question.** The memory question was, “In this study, you were asked to make decisions about and offer evaluations of a vaccine based on what?”

- Its efficacy at preventing infection and the duration of side effects it would cause.
- The cost of the vaccine and whether it works as an mRNA vaccine.
- The percentage of the population that has already taken the vaccine and its country of origin
- The number of miles one would drive to get the vaccine and the amount of time one would have to wait for the vaccine to achieve its full effect.

The accurate answer was “Its efficacy at preventing infection and the duration of side effects it would cause.”

**Analyses excluding those who failed to answer the memory question correctly.** We excluded 328 participants who were unable to answer the memory question correctly. Analyses were conducted on the remaining 896 participants.

We began by testing whether there was evidence that the evaluation task (more than the response task) placed those values into the decision sample, with predictable influence on vaccine decisions. This multi-level model included fixed effects of decision range (+1: high, -1: low) and task (+1: evaluation, -1: response) that characterized each participant’s condition assignment. Crucially, the interaction between these two variables was included as well. Furthermore, we included fixed effects of side effect and efficacy that described the vaccine’s standing on a particular decision trial. The side effect duration was centered by converting the values (1 to 10 days) into a -4.5 to +4.5 scale. The efficacy values were recoded by decision range condition to describe whether a particular decision trial’s vaccine was relatively efficacious or not: -2 (5% or 55%), -1 (15% or 65%), 0 (25% or 75%), 1 (35% or 85%), +2 (45% or 95%). To account for nonindependence across trials, we included participant as a random factor.

Unsurprisingly, we observed main effects of vaccine support,  $t(891) = 12.53, p < .001$ , decision range,  $t(891) = 18.47, p < .001$ , efficacy,  $t(43902) = 69.11, p < .001$ , and side-effect duration,  $t(43902) = -49.13, p < .001$ . In other words, people expressed more willingness to get a vaccine when: (a) they entered the experiment with more support for vaccines, (b) they made decisions about more efficacious vaccines, (c) they considered a vaccine that was more efficacious than the other vaccines they made decisions about, and (d) the vaccine promised

fewer days of side effects. All effects are quite sensible and quite well supported in the evidence. Though of central relevance, we observed a Decision Range X Task interaction,  $t(891) = 2.81$ ,  $p = .005$ . That we observed no main effect of task,  $t < 1$ , speaks to how the task manipulation operated symmetrically across the two decision ranges.

More specifically, when participants evaluated the other efficacy levels, their stated intention to receive the vaccine strongly depended on the decision range—i.e., whether they were making decisions about a set of quite efficacious vaccines (55% to 95% efficacy) or less efficacious vaccines (5% to 45% efficacy): 63.58% vs. 20.92 (42.66 percentage points). But when participants merely responded to by retyping the other efficacy levels, this gap was significantly reduced: 56.48% vs. 24.59% (31.89 percentage points). By the Evaluation Account, evaluating (more than merely responding to) values places those values in the decision sample, meaning that they are used in the decision sample for assessing the subjective efficacy magnitude. Evidence of this is seen in the greater spread between vaccine interest in the high and low decision range conditions among those who evaluated (versus merely retyped) the other attribute values.

Next, we moved to our test of whether the evaluation task (compared to the response take) changed how participants subjectively evaluated the efficacy of the trials in the post-trials evaluation measure. Given that participants made these judgments about 5 levels of efficacy (i.e., those that composed the vaccines about which they made decisions), we retained efficacy as a predictor but dropped side effect duration from the model (given that such durations were not manipulated on the post-trials evaluation measures). Once again, we observed main effects of vaccine support,  $t(891) = 11.51$ ,  $p < .001$ , decision range,  $t(891) = 22.86$ ,  $p < .001$ , and efficacy,  $t(3583) = 65.13$ ,  $p < .001$ .

But of central relevance, we again observed a Decision Range X Task interaction,  $t(891) = 2.78, p = .006$ . This supported our explanation that when participants had evaluated the other efficacy rates (i.e., those about which they did not make vaccine decisions), they showed a sizable gap in their evaluation of the high versus low range of efficacies: 60.84 vs. 24.89. But when participants had merely retyped the other efficacy rates, this post-trial evaluation gap shrank: 57.05 vs. 28.56.

Finally, we tested whether these post-trials efficacy evaluations statistically mediated our key interaction on the vaccine decisions. Toward that end, we added an *evaluation* variable to the original model. This value corresponded to how participants ultimately rated (on the post-trials evaluation measure) the specific efficacy level that described a particular vaccine trial. The proposed mediator (evaluation) significantly predicted participants' likelihood of accepting the vaccine on a specific trial,  $t(38189.23) = 61.56, p < .001$ . The Decision Range  $\times$  Task interaction remained significant,  $t(881.71) = 2.01, p = .045$ . This pattern of results is thus consistent with partial mediation, Sobel  $z = 2.77, p = .006$ .