Psychological Bulletin
2018, Vol. 144, No. 2, 140–176

0033-2909/18/$12.00 http://dx.doi.org/10.1037/bul0000115

A Meta-Analytic Review of Two Modes of Learning and the Description-Experience Gap

Dirk U. Wulff, Max Mergenthaler-Canseco, and Ralph Hertwig
Max Planck Institute for Human Development, Berlin, Germany

People can learn about the probabilistic consequences of their actions in two ways: One is by consulting descriptions of an action's consequences and probabilities (e.g., reading up on a medication's side effects). The other is by personally experiencing the probabilistic consequences of an action (e.g., beta testing software). In principle, people taking each route can reach analogous states of knowledge and consequently make analogous decisions. In the last dozen years, however, research has demonstrated systematic discrepancies between description- and experienced-based choices. This *description-experience gap* has been attributed to factors including reliance on a small set of experience, the impact of recency, and different weighting of probability information in the two decision types. In this meta-analysis focusing on studies using the sampling paradigm of decisions from experience, we evaluated these and other determinants of the decision–experience gap by reference to more than 70,000 choices made by more than 6,000 participants. We found, first, a robust description-experience gap but also a key moderator, namely, problem structure. Second, the largest determinant of the gap was reliance on small samples and the associated sampling error: free to terminate search, individuals explored too little to experience all possible outcomes. Third, the gap persisted when sampling error was basically eliminated, suggesting other determinants. Fourth, the occurrence of recency was contingent on decision makers' autonomy to terminate search, consistent with the notion of optional stopping. Finally, we found indications of different probability weighting in decisions from experience versus decisions from description when the problem structure involved a risky and a safe option.

Keywords: risky choice, decisions from experience, exploration–exploitation, recency, expected value maximization

One of the greatest inventions of all time is communication through written symbols. This uniquely human cultural adaptation enables people to take advantage of an information-sharing network of formidable power. By agreeing on the meaning of written and spoken symbols—the oldest known system of writing dates back to the Mesopotamian culture of the late fourth millennium BC (Schmandt-Besserat, 1996)—people benefit from the wisdom and knowledge accumulated by others through trial and error, lucky accidents, careful analysis, and the many other ways in which knowledge and intelligence is generated. The mind's ability to process written language liberates people from relying predominantly on direct experience of the world, which may sometimes be dangerous. Nevertheless, firsthand experience remains an important window onto the proximate environment. In fact, some languages (e.g., Turkish) require speakers to mark whether they have witnessed a particular event firsthand or learned about it through other means (Pinker, 2007).

This article was published Online First December 14, 2017.

Dirk U. Wulff, Max Mergenthaler-Canseco, and Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

Correspondence concerning this article should be addressed to Dirk U. Wulff, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, DE-14195 Berlin, Germany. E-mail: dirk.wulff@gmail.com

Assuming the knowledge encountered is structurally equivalent, do these two distinct modes of learning and intelligent adaptation (March & Olsen, 2010)—learning from descriptions in terms of written (or graphic) symbols versus learning from experience—result in similar outcomes? Folk wisdom is of two minds regarding the value of experience. It both trumpets the significance of direct experience—"experience is the best teacher"—and warns of its inadequacies—"experience is the teacher of fools," that is, of those unwilling to learn from accumulated knowledge. One field in which this question has received much attention in the last decade is behavioral decision research. What and how do people decide in stochastic worlds where options and outcomes are not certain but probabilistic?

Monetary Lotteries: Decision Science's Indispensable Fruit Fly

> There is no need to enlarge upon the importance of a realistic theory explaining how individuals choose among alternate courses of action when the consequences of their actions are incompletely known to them. . . . Risk and the human reactions to it have been called upon to explain everything from the purchase of chances in a "numbers" game to the capitalist structure of our economy. . . . (Arrow, p. 404)

Just as biologists have scrutinized the Drosophila (fruit fly) as a model organism, behavioral decision researchers have commonly used choice between monetary gambles (often called lotteries) as a model for the decisions people take. Admittedly, this fruit fly is a somewhat odd creature, divorced from any real content and

context. Yet it would be difficult to overstate the crucial role that gambles played in the development of the notion of mathematical expectation as well as normative and descriptive theories of choice. For instance, the fundamental principles of probability were first formulated in the mid-1600s in an exchange of letters between the French mathematicians Blaise Pascal and Pierre Fermat, who discussed various monetary gambling problems raised by the notorious gambler and man about town, Chevalier de Méré. This exchange gave rise to the concept of mathematical expectation (Hacking, 1975). Another gambling problem, the St. Petersburg lottery, and the seemingly odd behavior it produced (i.e., behavior deviating from mathematical expectation) gave rise to Daniel Bernoulli's (1738/1954) revision of expected-value theory, today known as *expected utility theory* and the backbone of classic economic theory. Monetary gambles were also enlisted to show that people's choices are often at odds with expected utility theory (e.g., Allais, 1953; Ellsberg, 1961) and to develop descriptive alternatives, such as prospect theory (Kahneman & Tversky, 1979). Last but not least, in his groundbreaking book *Foundations of Statistics*, Savage (1954/1972) suggested that he believed gambling decisions—such as the six-egg omelet problem—are representative of all decisions people take. From this perspective, countless real-life decisions have the same structural properties as gambles, namely, acts (choices), consequences (outcomes), and the consequences' known or unknown probabilities.

Many researchers in both psychology and economics have grown accustomed to presenting respondents with one particular representation of this paradigmatic choice, namely, lotteries in which consequences (possible outcomes) and probabilities are explicitly stated. That is, respondents are told about the stochastic texture of their choice environment through written or graphic symbols, and their knowledge of the options is typically complete. Some researchers have referred to this kind of decision task as "static" (Edwards, 1962) and noted that:

> [w]hen a static decision task is used, the decision maker does not have to learn from past experience with the outcomes of previous decisions. . . . This feature of the static decision task becomes a problem when generalizing results to the many day-to-day decisions that repeatedly confront individuals, since explicit information concerning outcome probabilities is frequently not available and must be learned from previous experience. (Busemeyer, 1982, p. 176)

Indeed, in everyday life, people can rarely consult explicit descriptions of probability distributions (with a few exceptions, such as leaflets listing the probabilities of side effects of medications, or weather forecasts stating probabilities of precipitation; e.g., Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005). When people decide whether to start a business or contemplate the success of a first date, there are no written records of risks to consult. Instead, they need to rely on their experience—if existent—with these options, and make *decisions from experience* rather than *decisions from description* (Hertwig, Barron, Weber, & Erev, 2004). This distinction has raised a new and deceptively simple question: To what extent do these two modes of learning about the world—which we understand as poles on a continuum—result in similar or systematically different choices? This question has received much attention since a set of three articles in the early 2,000s demonstrated a systematic discrepancy in description- and experienced-based choices: the description-experience gap (Barron & Erev, 2003; Hertwig et al., 2004; Weber, Shafir, & Blais, 2004). This meta-analysis synthesizes the now extensive corpus of studies addressing this gap.

To conclude, despite their abstract structure, monetary gambles have played a central role in the development of probability theory, normative and descriptive theories of individual choice. They are also *the* measurement tool for key constructs in economic and psychological theorizing, such as risk aversion, loss aversion, and ambiguity aversion (e.g., Holt & Laury, 2002). Before we turn to the description-experience gap, let us emphasize that the experimental study of decisions from experience is, of course, nothing new. Ward Edwards and other godfathers of modern decision science engaged in it (see Luce & Suppes, 1965, for a review) as did early probability learning researchers (see Lee, 1971, for a review). Although behavioral decision researchers began to turn away from the transients of learning in the 1970s (with some notable exceptions; e.g., Busemeyer, 1982, 1985), concerns for experience-based decision making remained alive in areas of economics (e.g., reinforcement learning in experimental games; e.g., Erev & Roth, 1998) and operation research (see the literature on multiarmed bandit problems; Sutton & Barto, 1998). The novel contribution of research on the description-experience gap has been to systematically pit decisions from experience against decisions from description, commonly using monetary lotteries but increasingly going beyond this approach (see General Discussion section). This research has raised new questions, including the extent to which apparent regularities in human choice generalize from description to experience (e.g., loss aversion, nonlinear probability weighting, the fourfold pattern of risk attitude; see Kahneman & Tversky, 1979).

Decisions From Description and Decisions From Experience

In description-based studies, individuals are presented with descriptions of payoff distributions (typically written or graphic representations) and asked to choose between, for instance, the following options A and B:

A: \$4 with a probability of 80%, and 0 otherwise
or
B: \$3 for sure.

In this choice, individuals receive *complete* information about all possible outcomes (4, 3, and 0) and their respective quantitative likelihoods (.8, 1, and .2). In the parlance of behavioral decision science, they are thus making choices under risk (e.g., Knight, 1921; Luce & Raiffa, 1957/2012). Of course, descriptions may be incomplete; in this case, they do not represent choices under risk but choices under uncertainty or ambiguity. In Ellsberg's (1961) examination of choices under ambiguity, the description of some lottery options is incomplete to the extent that all possible events are stated but not their probabilities (see also Trautmann & van de Kuilen, 2016).

In experience-based studies, in contrast, individuals are initially in a state of ignorance about the properties of the payoff distributions. However, they have the opportunity to explore the distributions to develop an understanding of the underlying structure. A typical study presents individuals with two (or more) boxes rep-

resenting unknown payoff distributions. In order to learn about the possible outcomes and frequencies, participants can draw random samples from each of the payoff distributions. This exploration process is typically under the participants' own control: They decide how long to explore, which option to explore, and when to switch between options. Figure 1 shows the three related but distinct experimental paradigms that have been employed to study decisions from experience (Hertwig & Erev, 2009); there are also many hybrids, in which descriptive and experiential information is combined (e.g., Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Erev, Glozman, & Hertwig, 2008; Lejarraga & Müller-Trede, 2016). In the *sampling paradigm* (A), participants may explore the options for as long as they like before making one final incentivized choice. Except for opportunity costs (time) and cognitive and motor effort, no search costs are incurred during sampling. In the *partial-feedback paradigm* (B), each single draw from the payoff distributions represents both payoff and information. As a consequence, participants seeking to boost their income in the long run face a tension between efforts to learn more about the world (exploration) and efforts to take advantage of what is already known (exploitation)—that is, the exploration–exploitation tradeoff (Cohen, McClure, & Yu, 2007). In the *full-feedback paradigm* (C), this tradeoff is suspended—not by separating exploration and exploitation, as in the sampling paradigm, but by providing full information: participants learn about both the obtained and the forgone payoff (the outcome they would have obtained had they selected the other option).

In all three paradigms, participants typically start out with no knowledge of the payoff distributions; they gain some knowledge through experiencing the outcomes sampled, and thus make decisions from experience. Crucially, the probabilities of the experienced outcomes can only be estimated, rather than known with certainty (to use Knight's, 1921, terminology, they are *statistical probabilities rather than a priori probabilities*). Consequently, participants' knowledge of the outcomes' objective probabilities is *vague* or *imprecise* (see Budescu & Wallsten, 1987). Although precision increases with sample size, in any finite number of trials, no matter how large, it is impossible to identify the true probabilities with certainty. A similar dynamic holds for the outcome space. Even when very large samples are drawn, knowledge of the outcome space may remain incomplete (e.g., because a rare event has not yet been encountered) and is never certain.

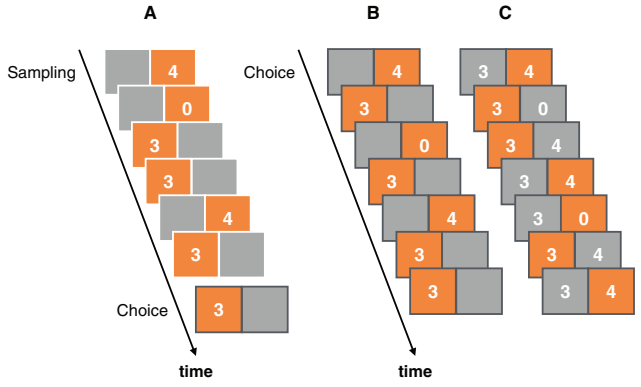

Figure 1. Illustration of the three main paradigms of decisions from experience: the sampling paradigm (A), the partial-feedback paradigm (B), and the full-feedback paradigm (C). See the online article for the color version of this figure.

The Description-Experience Gap

As Figure 1 shows, participants deal with the same objective payoff distributions across the three paradigms. However, the payoff distributions experienced may deviate systematically from the objective (described) distributions, depending on participants' exploration (and exploitation) of the options. As a result, experienced-based choices may deviate systematically from description-based choices. Indeed, a systematic description-experience gap in choice behavior has been observed across all three paradigms. For illustration, consider the fourfold pattern of risk attitudes, a classic finding in choice under risk (Tversky & Fox, 1995; Tversky & Kahneman, 1992). The fourfold pattern refers to the phenomenon that people are generally risk averse when the *stated* probability of winning is high but risk seeking when it is low (as when buying lotteries) and risk averse when the stated probability of losing is low (as when buying insurance) but risk seeking when it is high. Table 1 illustrates the classic fourfold pattern in decisions from description and its reversal in decisions from experience, using results from Hertwig et al. (2004; see also Hertwig, 2012). Note that the fourfold pattern rests on the definition of risk that is standard in many models of decision making under risk and uncertainty in economics and psychology, namely, as the variance of outcomes around the option's expected value.

The standard explanation for the fourfold pattern is in terms of nonlinear probability weighting (Tversky & Kahneman, 1992). Probability weighting refers to the assumption that the value of an outcome is multiplied by a decision weight that captures the subjective impact of that outcome on choice and that can deviate from the outcome's stated probability. In Table 1, the choice proportions for description-based choices in the gain and loss domain are consistent with the assumption in prospect theory that unlikely (i.e., rare) events are overweighted (Kahneman & Tversky, 1979). For instance, in the loss domain, the majority's preference for the risky option (−4, .8, 0 otherwise) over the safe option (−3) is consistent with overweighting the impact of the relatively unlikely (.2) but desirable event 0; in the gain domain, the same unlikely outcome is undesirable, and overweighting its impact makes the risky option less attractive, thus explaining why the majority choice is risk averse. In decisions from experience, both of these majority preferences are reversed. The overall pattern can be summarized as follows: In decisions from experience, people behave *as if* rare events have less impact than they deserve according to their objective probabilities, whereas in decisions from description, people behave *as if* rare events have more impact than they deserve (Hertwig & Erev, 2009). To what extent this implicit portrayal of the shape of the probability weighting function and its reversal (inferred from the choices observed in decisions from description and experience) is appropriate will be discussed later.

What causes this description-experience gap in choice? A number of determinants have been suggested and examined. This meta-analysis will focus on the sampling paradigm (see Figure 1), which has been the focus of most experimental work to date. Three main determinants of the description-experience gap have been

Table 1
The Fourfold Pattern of Risk Attitudes in Decisions from Description and it's Reversal in Decisions from Experience (Based on Hertwig et al., 2004)

	Decisions from description		Decisions from experience	
Probability	Gain domain	Loss domain	Gain domain	Loss domain
Low	32, .1[a] vs. 3, 1.0 Rare event: 32, .1 Risk seeking 48%[b]	−32, .1 vs. −3, 1.0 Rare event: −32, .1 Risk averse 36%	32, .1 vs. 3, 1.0 Rare event: 32, .1 Risk averse 20%	−32, .1 vs. −3, 1.0 Rare event: −32, .1 Risk seeking 72%
High	4, .8 vs. 3, 1.0 Rare event: 0, .2 Risk averse 36%	−4, .8 vs. −3, 1.0 Rare event: 0, .2 Risk seeking 72%	4, .8 vs. 3, 1.0 Rare event: 0, .2 Risk seeking 88%	−4, .8 vs. −3, 1.0 Rare event: 0, .2 Risk averse 44%

[a] For the sake of brevity, the alternative outcome (0 otherwise) has been omitted for all risky options. [b] Proportion of risky choices. In past studies, this proportion has been found to be greater than 50% (e.g., Tversky and Kahneman (1992)).

studied within this paradigm (for others, see Hertwig, 2016; Rakow & Newell, 2010): (a) reliance on small samples, (b) recency, and (c) reversal of probability weighting. Work on reliance on small samples suggests that the gap results partly from the sampling error associated with limited exploration: Decisions from description differ systematically from descriptions from experience because the stated and the experienced options can diverge systematically. Research on recency suggests that explorers give more weight to outcomes that occurred more recently in their flow of experience than to earlier outcomes. In other words, even if the experienced frequencies veridically track the stated probabilities, recency can systematically misalign them (a small sample of *recent* events introduces a systematic sampling error). Finally, findings on reversed probability weighting suggest, as outlined in our account of the fourfold pattern of risk attitudes, that decisions from description and decisions from experience evoke distinct probability-weighting patterns and, in particular, the overweighting (description) and underweighting (experience) of rare events.

In what follows, we draw on the extensive body of data collected in the sampling paradigm to meta-analytically investigate, first, the size and robustness of the description-experience gap and, second, the impact of its three potential determinants.

Database and Literature Search History

Our aim was to collect all data sets (published or unpublished) that employed the sampling paradigm and were available by December 2015. First, we identified all articles in the Thomson Reuters *Web of Science* that cited Hertwig et al. (2004), the article that is often referenced as having established the sampling paradigm. Of the 304 articles identified, 152 reported original data and, among those, 77 required participants to sequentially choose between experienced options. This was the criterion for including an article in our consideration set. Second, to ensure the data sets included in the analysis were sufficiently comparable, we applied six inclusion criteria: (a) *Payoff distributions*: Participants decided between two options, each with one or, at most, five possible outcomes. (b) *Game against nature*: The problem involved individual decision making; articles dealing with strategic concerns about other players within the sampling paradigm were excluded (e.g., the experimental condition of Phillips, Hertwig, Kareev, & Avrahami, 2014). (c) *Initial ignorance*: No information about the payoff distributions was provided at the outset. (d) *Active search*: Participants engaged in active search, that is, their search decisions triggered sampling from the payoff distributions. (e) *Costless search*: Search incurred no monetary costs; it involved only opportunity costs and cognitive and motor effort. This criterion was chosen to exclude studies employing the full- and partial-feedback paradigms (see Figure 1). (f) *Independent and identically distributed random variables*: Sampling from the payoff distribution was random or pseudorandom (e.g., Camilleri & Newell, 2011a). The application of these criteria led to set of 29 articles.

Third, to make sure that we did not overlook any relevant articles, we screened all articles identified by *Web of Science* and *Google Scholar* as citing any of the three seminal articles on the description-experience gap (i.e., Barron & Erev, 2003; Hertwig et al., 2004; Weber et al., 2004) or the review article by Hertwig and Erev (2009). In addition, we conducted a keyword search in *Web of Science* using the terms "decisions from experience" and "description-experience gap." This resulted in a total of 1,916 further hits. However, no additional articles matching our criteria were identified.

Next, we contacted the authors of the 29 articles identified for inclusion in the analysis to request their original data. In 25 cases, the authors made the raw data available. In three cases, the experimental protocol did not record the participants' sampled sequences of outcomes (Barron & Ursino, 2013; Hilbig & Glöckner, 2011; Weber et al., 2004). In one case (Erev, Glozman, & Hertwig, 2008), the data were lost. Finally, we added an unpublished replication of Hertwig et al. (2004) conducted by Wulff and Hertwig (2012) with two online samples acquired through Amazon Mechanical Turk.

In December 2016, we extended our literature search by screening all articles published in 2016 that the *Web of Science* identified as citing one of the four key articles listed above. Of the 82 articles identified, 58 reported original data, 30 required participants to sequentially choose between experienced options and, of those, five articles met the six inclusion criteria listed above. Three articles were already included in our analysis as advance online publications. The authors of the remaining two articles provided us with the raw data.

In addition, we contacted the first authors of the now 28 articles included in the meta-analysis and asked them whether they had

any unpublished data relevant for the purposes of our meta-analysis or knew of any relevant unpublished data by another author (collected prior to December 2015). One dataset was suggested (Noguchi & Hills, 2016), but it did not match our inclusion criteria. For additional details on the process of contacting the authors, see van den Bos, Jenny, and Wulff (2014).

This extensive search produced a final pool of 80 data sets (some representing different studies or conditions within one publication) in which a total of 4,400 participants made 45,239 decisions from experience. In addition, 2,208 participants made 31,353 decisions from description (some data sets conducted only an experience condition). Details of the data sets and their origins are provided in Appendix A. The specific implementation of the sampling process varied across the 80 experience-based data sets: 55 implemented *autonomous sampling*, that is, participants decided how much they wanted to sample, and samples were completely random independent and identically distributed (iid). Four data sets implemented matched-autonomous sampling (henceforth *matched sampling*), that is, participants terminated search of their own accord, but a pseudorandom process aimed to "match" experienced frequencies and described probabilities. Finally, 21 data sets implemented *regulated sampling*, that is, the experimenter predetermined the number of samples to be observed (e.g., $N = 20$), and samples were generated either randomly or pseudorandomly. Table 2 offers more detail on the data sets, including information on the domain from which decision problems were selected.

Finally, for the purpose of reviewing studies of probability weighting, we conducted an independent search—the reason being that our original search criteria would have excluded some pertinent studies that used modifications of the sampling paradigm (Abdellaoui, L'Haridon, & Paraschiv, 2011; Jarvstad, Hahn, Rushton, & Warren, 2013; Kemel & Travers, 2016; Zeigenfuse, Pleskac, & Liu, 2014). Specifically, we conducted a keyword search in *Web of Science* using the terms "decisions from experience" AND "probability weighting," "decisions from experience" AND "prospect theory," "description-experience gap" AND "probability weighting," and "description-experience gap" AND "prospect theory." This search produced a total of 56 hits. We screened all articles to identify those that actually measured the probability weighting parameters assumed in cumulative prospect theory. This search was updated in December 2016. We found nine relevant articles. To these, we added three articles that were known to us but not (yet) listed in *Web of Science* (Glöckner et al., 2016; Kellen, Pachur, & Hertwig, 2016; Markant, Pleskac, Diederich, Pachur, & Hertwig, 2015).

Using the raw, trial-level data, we analyzed both commonly studied effects (e.g., the description-experience gap) as well as less commonly studied effects (e.g., the description-experience gap for equivalent experience). To investigate the possibility that studies finding nonsignificant results for the commonly studied effects were less likely to be accepted for publication, we used funnel plots. Funnel plots are scatterplots of effect sizes in primary studies and their standard errors; asymmetric funnel plots may indicate publication bias (Light, Singer, & Willet, 1994). We used Egger's linear regression method to test for funnel plot asymmetry and the "trim and fill" method to impute suspected missing studies until the studies were symmetrically distributed around the pooled effect size (Duval & Tweedie, 2000). These analyses were carried out using the R package *metafor* (Viechtbauer, 2010). To further reduce the risk of publication bias, we also asked authors to provide us with any unpublished data sets, as described before. Finally, let us report our impression that because the existence of a description-experience gap has not remained unchallenged, both studies that observed the gap and those that observed conflicting evidence found their way into the journals.

The Description-Experience Gap: How Large and How Robust is It?

In order to address this question, we first describe two ways in which researchers have operationalized the description-experience gap. A very simple definition would treat any (significant) discrepancy between choice proportions in the description and the experience condition as evidence for a description-experience gap. However, this definition would not permit any predictions to be made about systematic differences (i.e., in which of the two conditions the choice proportions will be higher/lower). To render possible directional predictions, Barron and Erev (2003) and Hertwig et al. (2004) capitalized on the assumed as-if weighting of rare events in combination with the events' desirability in decisions from description (Kahneman & Tversky, 1979). Specifically, if the rare event (defined as .2 or below;[1] see Hertwig et al., 2004, Footnote 2) was desirable (i.e., represented the largest positive or the least negative outcome), then any time the option with the rare event was chosen more often in the description condition than in the experience condition was scored as an instance of the gap. This scoring follows from the assumed as-if over (description) versus underweighting (experience) of rare events. For undesirable rare events, the directionality of scoring was reversed. Most studies have employed this *discrete* way to operationalize the description-experience gap.

Another approach to operationalizing the description-experience gap has been to count how many of individuals' *actual* choices in the description and experience condition, respectively, are consistent with the *predicted* choices, based on cumulative prospect theory's (CPT) parameters (commonly using those derived by Tversky & Kahneman, 1992). Note that these parameters, derived from stated probabilities, embody overweighting of rare events; therefore, choices consistent with the predictions of CPT indicate a tendency to overweight rare events. When this definition is applied, a description-experience gap emerges when systematically fewer experienced-based than description-based choices are correctly predicted. A number of studies have employed this definition (e.g., Camilleri & Newell, 2009a; Fox & Hadar, 2006; Ungemach, Chater, & Stewart, 2009).

How different are the discrete and the CPT-based operationalizations of the description-experience gap? In the present pool of experienced-based decisions, we found that they result in identical predictions in 67% of cases. Thus, in order to investigate the robustness of the gap across these two most fre-

[1] Hertwig et al.'s (2004) working definition of events as rare ($p \leq .2$) and common ($p > .2$) was practical but admittedly arbitrary. Of course, there is no evidence that individuals treat events of $p = .19$ any differently than they do events of $p = .21$. In fact, whether the weighting of probabilities in decisions from experience exhibits nonlinearities that could justify a categorization into rare and nonrare events is entirely unclear. Moreover, as we will show in the section on sampling error (see Figure 5), the notion of rare events is not *necessary* to account for the as-if underweighting observed in decisions from experience (as has also been demonstrated by Ludvig & Spetch, 2011).

Table 2
Characteristics of the Experienced-Based and Description-Based Data Sets

	Data sets			Decision domains				
	Data sets (*N*)	*N*	Choices (*N*)	Gain (*N*)	Loss (*N*)	Mixed (*N*)	Certain	Binary
Experience	80	4,400	45,239	26,514	13,216	5,509	33%	86%
Autonomous	55	2,643	40,246	23,067	11,962	5,217	27%	84%
Matched	4	226	977	621	356	—	100%	100%
Regulated	21	1,531	4,016	2,826	898	292	84%	100%
Description	48	2,208	31,353	18,182	7,859	5,312	33%	99%

Note. *N* refers to number of participants. Certain indicates the percentage of cases in which one of the options offered a certain outcome. Binary indicates the percentage of cases in which both options had at most two outcomes.

quently used operationalizations in the literature, we will report the results for both.

How large and how robust is the description-experience gap? In addressing this question, we included all data sets that had a description condition and an analogous experience condition, that implemented autonomous sampling, and that did not implement an intervention designed to reduce the gap (a subset of studies to which we return later). As Figure 2 shows, we found a gap

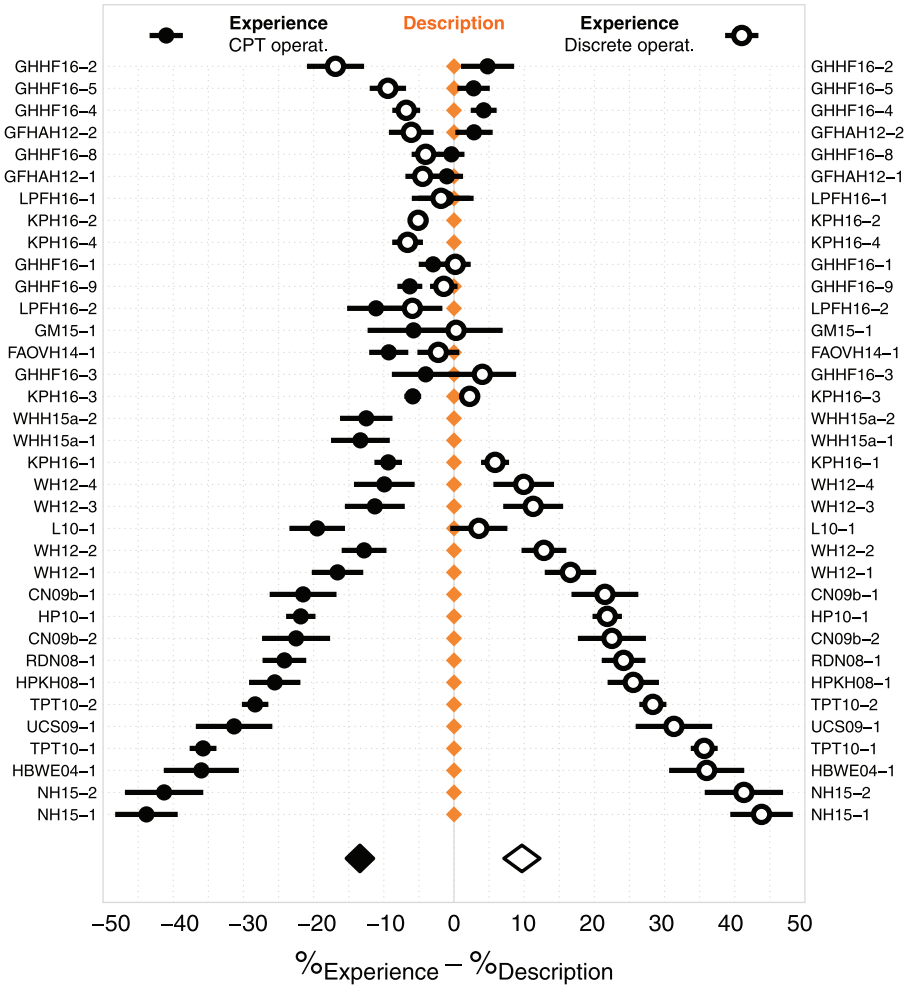

Figure 2. The description-experience gap. Difference in proportion of choices between the description and experience conditions, based on the two operationalizations of the gap (see text and Footnote 2). Error bars represent the standard error of the mean. Diamonds represent aggregate estimates and standard errors based on random effects meta-analysis. See Table A1 in Appendix A for a key to the study abbreviations. See the online article for the color version of this figure.

between description and experience consistent with a discrete as-if underweighting of rare events in experience relative to description in 21 of the 33 data sets.[2] Consistent with a CPT-based operationalization of the description-experience gap, the proportions of correctly predicted choices were smaller in the experience than in the description condition in 29 of the 35 data sets.

As estimated from a random effects meta-analysis,[3] description and experience conditions differed by, on average, 9.7 percentage points ($z = 3.49$, $p < .001$) in the discrete operationalization and by −13.4 percentage points in the CPT-based operationalization ($z = -6.27$, $p < .001$). The odds of choosing the option with the (as-if underweighted) rare event or the option inconsistent with the (as-if CPT overweighted) rare event were 1.43 and 1.77 times greater, respectively, in the experience than in the description condition. We detected a small funnel asymmetry for both discrete underweighting ($z = 3.41$, $p = 3.41$) and the CPT-based operationalization ($z = -6.56$, $p < .001$). However, in neither case were the results affected when this publication bias was taken into account (see Appendix B). To conclude, these results suggest that, irrespective of how the gap is measured, the description-experience gap is robust and substantial.

As a first step toward analyzing factors that increase or attenuate the description-experience gap, we next investigated the following variables: (a) the structure of the decision problem (i.e., choice between a risky and a certain option vs. two risky options, with "risky" being defined as an option with more than one possible outcome); (b) the probability of the rarest event; and (c) the domain of the decision problem (gain vs. loss vs. mixed). We found three key results (see Table 3). First, for both operationalizations, the size of the gap was *strongly* affected by the problem structure. When a choice involved a risky and a safe option, the gap was large, namely about 20 percentage points; see Figure 3 and Appendix C). When a choice involved two risky options, however, the gap was nearly eliminated[4] under discrete underweighting and substantially reduced under the CPT-based operationalization (see Figure 3 and Table 3). Note that only 28% of the choices analyzed involved a risky and a safe option, although most studies employed this type of choice. Two recent studies (Glöckner, Hilbig, Henninger, & Fiedler, 2016, and Kellen, Pachur, & Hertwig, 2016), both of which aimed to analyze the probability-weighting pattern in experienced-based choice, contributed 56% of the choices involving two risky options. Second, for both operationalizations, the rarer the outcome, the larger the gap (but only in options involving one risky and one certain option; Figure 3). Third, the domain of the decision problem moderated the size of the gap under the CPT-based operationalization, with loss and mixed problems leading to larger gap sizes than gain problems, but not under discrete underweighting. Finally, we tested for a moderating role of financial incentives (i.e., whether participants received a bonus contingent on their choices) and practice effects (i.e., how early or late in the sequence of decision problems a particular choice was made). We found evidence for the latter but not the former. That is, the gap was reduced in later parts of the sequence (see Table 3).

To conclude, we found that two structural properties of the payoff distributions strongly affect the description-experience gap: the structure of the problem and the probability of the rarest event. Both properties also moderate the description-experience gap in the partial-feedback paradigm, and a preliminary analysis suggests that the magnitude of the gap for choices between risky and safe options is comparable in the sampling and partial-feedback paradigms (see Appendix C). In addition, we found evidence that both the domain of the decision problem and practice effects play a moderating role. We next turn to potential determinants of the gap that do not reside in the environment but are located within the mind of the decision maker.

Table 3
Moderators of the Description-Experience Gap

Moderator	DU	CPT
Structure	**Log-odds = .99, $p < .001$**	**Log-odds = −.64, $p = .037$**
Rarity	**Log-odds = −2.35, $p < .001$**	**Log-odds = 1.55, $p < .001$**
Domain	Log-odds = −.00, $p = .950$	**Log-odds = .21, $p < .001$**
Incentives	Log-odds = −.02, $p = .876$	Log-odds = −.02, $p = .948$
Order[a]	**Log-odds = −.10, $p < .001$**	**Log-odds = .06, $p = .004$**

Note. Effects are calculated on the basis of interactions of the respective variables with a dummy variable (1 = experience, 0 = description) within a mixed effects model controlling for study and participant effects.
[a] Because order information was available for only a small subset of studies, its effect on the gap was tested in a separate regression.

Does Sample Size and Sampling Error "Explain" the Description-Experience Gap?

Decisions from experience and decisions from description present payoff distributions that are, in theory, identical but can, in practice, be quite different: In decisions from experience, people's perception of the distributions is filtered through their sampled experience. Draws from the payoff distributions are commonly implemented as sampling of independently and identically distributed (iid) random variables (but see, e.g., Plonsky, Teodorescu, & Erev, 2015). Depending on the number of draws and the property of the distributions (e.g., the skewness of the payoff distribution), the experienced frequencies will be a better or worse proxy of the true probabilities. Consequently, the decision problem that a person in the experience condition faces may be systematically different from the analogous problem in the description condition, especially if that person's sampling is frugal. Hertwig et al. (2004) observed, on average, a median of 15 draws from both distributions. How representative was this early finding?

[2] Note that because discrete underweighting is applicable only to binary problems, the Wulff, Hills, and Hertwig (2015a) data were considered only for the CPT-based operationalization.
[3] Wherever possible, significance tests were based on subject- or trial-level linear mixed-effects models calculated using the R packages *lme4* and *lmerTest* (R Core Team, 2015).
[4] Is the CPT-based operationalization more robust toward the structure of the decision problem? One reason for this being likely to be the case is that environments with more than one risky option often contain a second small probability event (e.g., $4 with probability .2 and $3 with probability .25). In such cases, discrete underweighting considers only the rarer of the two events ($4 with probability .2), whereas the CPT-based operationalization can take both into account as a function of their rarity. The CPT-based operationalization may thus be considered more appropriate for choices with two risky options.

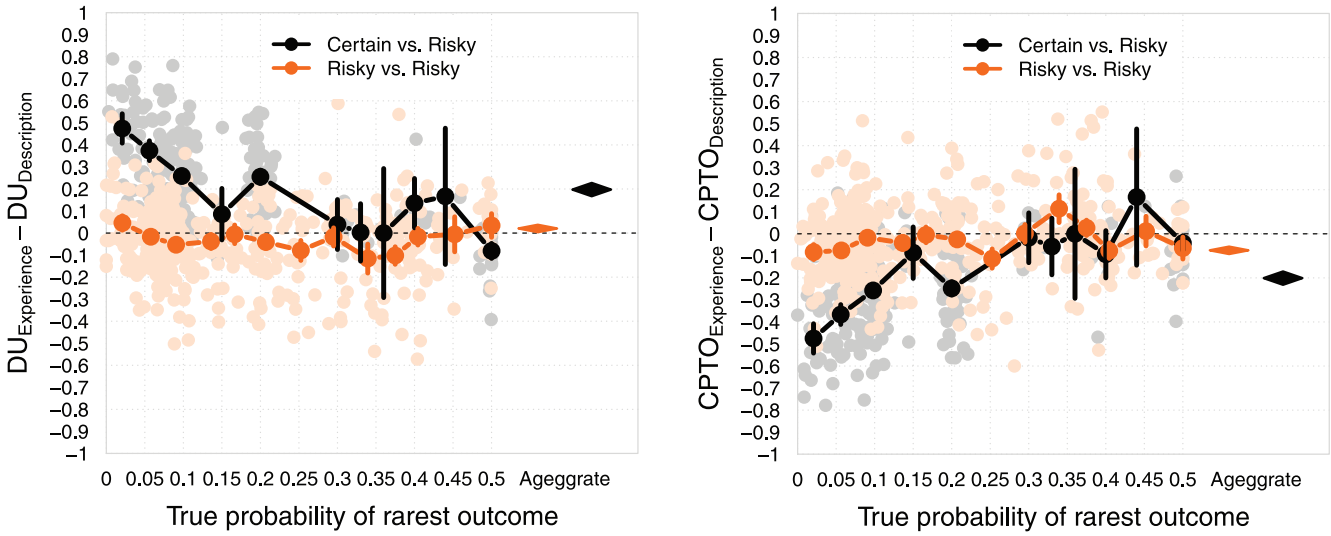

Figure 3. The structure of the decision problem and rarity. The two significant moderators of the description-experience gap as defined in terms of discrete underweighting (DU; left) and the CPT-based operationalization (CPTO; right) for autonomous sampling paradigms that did not implement an intervention designed to reduce the gap. Points in the background represent the description-experience gap for decision problems that required a choice between a certain and a risky option (gray) versus a choice between two risky options (orange). The lines in the foreground show the aggregate results for bins of size .04 in terms of the true probability of the rarest event. Error bars represent 95% confidence intervals. Diamonds represent aggregate estimates and standard errors based on random effects meta-analysis. See the online article for the color version of this figure.

Sample Size and Sampling Error

Figure 4 plots the sample size (number of draws) from the two payoff distributions for all 55 studies that implemented autonomous sampling. The median sample size across all 40,246 trials (problems × participants) is 20; for all choices involving a risky and a safe option (10,712 trials) it is 14; and for all choices involving two risky options (29,534 trials) it is 22. One benchmark against which sample size can be evaluated is the experience (and, by extension, awareness) of the complete outcome space.[5] Measured against this criterion, the sample sizes are relatively small. In about one third of the trials (36%) in the data sets (in Figure 4), at least one of the outcomes was not experienced. Because most options included a maximum of two outcomes, not having experienced one of them implies that the evaluation of the payoff distribution is, at least, incomplete. Small samples carry the risk not only that an event is not encountered but also that the relative frequency of an experienced event is misrepresented. Across all trials, we found that the experienced frequencies accurately tracked the true probabilities in only 1% of cases.

Sampling error not only enters noise into the experienced frequency of occurrence but can systematically distort it. Strictly speaking, whenever individuals do not engage in infinite sampling, the majority of them will experience rare outcomes less often than expected. To appreciate this fact, let us turn to the binomial distribution. It describes the number of successes of a Bernoulli trial with success probability p over N attempts. The skewness sk of the binomial distribution is calculated as follows:

$$sk = \frac{1 - 2p}{\sqrt{Np(1-p)}}.$$

This term will be positive (i.e., right-skewed) for all $p < .5$ and it will increase with smaller ps and smaller Ns. That is, the sampling distribution is particularly skewed for small sample sizes and small probabilities.[6] Figure 5 illustrates this regularity for all true probabilities across the 40,246 autonomous sampling trials analyzed here. Specifically, the medians of experienced relative frequencies are located below the identity line for small probabilities, tend to be clustered around the identity line for medium probabilities, and are located above the identity line for large probabilities. The plots of the marginal distributions in Figure 5 show that, relative to the decision problems' objective probabilities, the experienced relative frequencies are systematically shifted toward the boundaries (0 and 1).

[5] How many draws should an individual optimally take before making a choice? As shown in Ostwald et al. (2015; see also Vul, Goodman, Griffiths, & Tenenbaum, 2014), any principled answer hinges on numerous and often strong assumptions. Specifically, optimal sample size depends on (a) the subjective cost of sampling, (b) the subjective value of making a correct choice, and (c) the individual's prior beliefs about the nature of the decision problem (e.g., the range of possible outcomes).

[6] Research on risky choice generally relies on skewed distributions, such that medium-sized outcomes within the outcome space occur with high probability, whereas extreme outcomes occur with low probability. Using such skewed distributions has two advantages. First, they permit the construction of thorny problems, in which a person must trade off the likelihood of receiving the desired return against the magnitude of the return. Second, many natural environments involve such skewed distributions, where extremely positive outcomes (e.g., a lottery win) or extremely negative outcomes (e.g., a fatal car accident) are very rare (Pleskac & Hertwig, 2014).

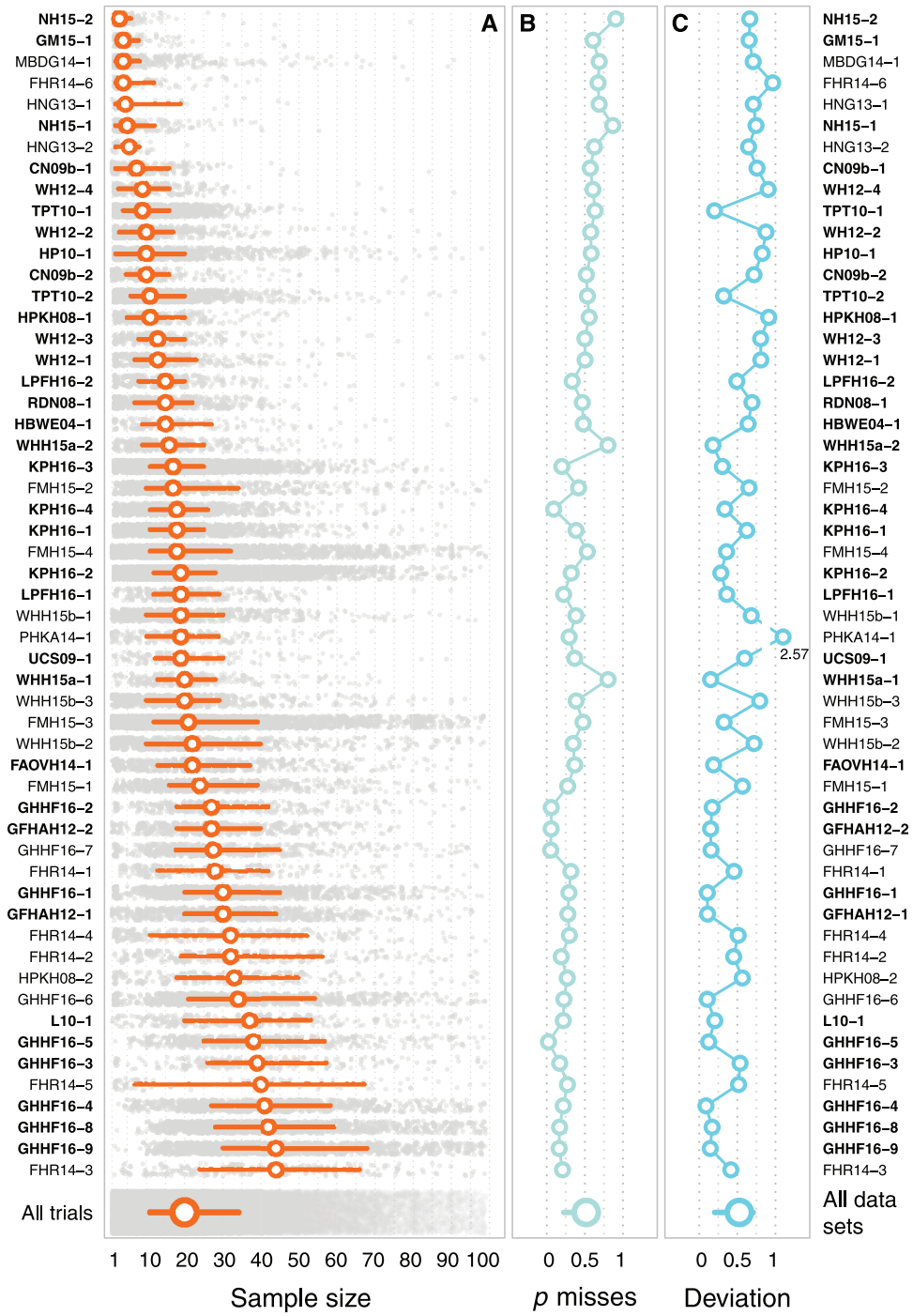

Figure 4. Sample size. Panel A plots the sample sizes across all 55 autonomous sampling data sets. The orange circles represent the median sample size for all trials within a data set. Panel B plots for each data set the proportion of trials in which at least one outcome was not experienced (i.e., misses). Panel C plots for each data set the magnitude of the sampling error, measured as the average absolute deviation from (and normalized by) an option's expected value (computed only for options with more than one outcome). Error bars represent the interquartile range (1st–3rd quartile). Labels in boldface are included in the meta-analysis on the description-experience gap. See Table A1 in Appendix A for a key to the study abbreviations. See the online article for the color version of this figure.

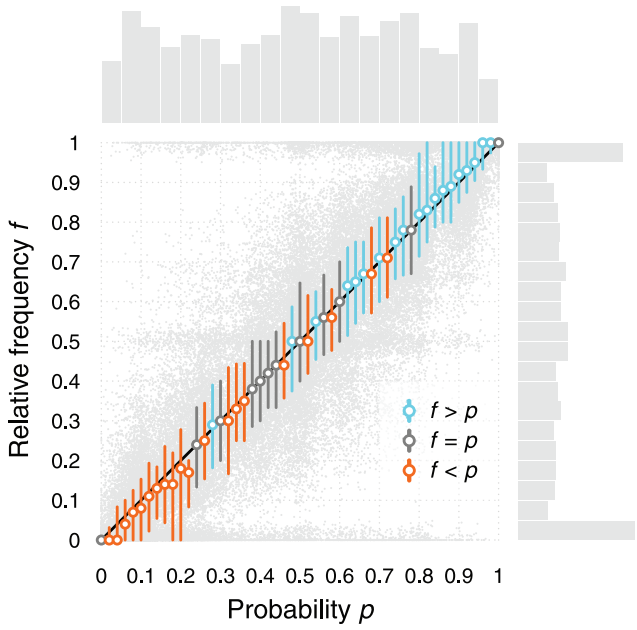

Figure 5. Small samples, sampling error, and rare events. This figure plots the true probability p versus the experienced relative frequency f for all autonomous sampling trials in our analysis (excluding options with $p = 1.0$). The points in the background represent the individual trials. The circles and lines in the foreground (in blue, gray, or orange) represent the median experienced probabilities for each unique true probability and the respective interquartile range. The bar graphs at the top and on the right show the marginal distribution of the objective probabilities and of the experienced relative frequencies, respectively. See the online article for the color version of this figure.

Choice Difficulty and Value Maximization

Which psychological factors shape individuals' sampling and explain the generally frugal nature of search? Hertwig and Pleskac (2010) proposed an amplification effect on choice difficulty as a possible reason for frugal search. Specifically, they derived analytically that small samples amplify the difference between the average earnings associated with payoff distributions, thus making choice options more distinct and choice easier. One interesting consequence of this possible dynamic is that choice may be more consistent with the maximization of expected earnings in experienced-based environments than in decisions from description. In the latter, expected earnings refer to the options' expected values; in the former, they refer to options' *experienced* mean returns. Indeed, it is well known that in lotteries with stated probabilities, the lower the choice difficulty (i.e., the larger the difference between the options' expected returns), the more likely choices are to agree with expected value maximization (see, e.g., Figure 8 in Brandstätter, Gigerenzer, & Hertwig, 2006).

To examine this possibility, we calculated for each (autonomous) sampling sequence the difference in the experienced mean returns of both options divided by the larger of the two values. We found a median difference of about 40%, relative to 19% in decisions from description (in the latter, we defined the difference between the options based on their expected values). That is, in decisions from experience, individuals encountered a difference between options that was about two times larger than in decisions from description, consistent with an amplification effect (Hertwig & Pleskac, 2010). Furthermore, we found experienced-based choices to be much more consistent with maximization of mean returns than with maximization of expected values in description-based choices. As Figure 6 shows, in decisions from description, a median of 55% of choices maximized expected value; in decisions from experience, 66% and 89% maximized the experienced mean return (depending on whether a sequence did or did not include all possible outcomes).

Moderators of Search

The interplay of the amplification effect and choice difficulty is, however, only one of several factors potentially shaping sampling. Table 4 summarizes determinants of search that have been identified in the literature to date. Some relate to the motivational, affective, and cognitive internal state of the sampler, such as the impact of incentives, vigilant emotion (fear), working memory capacity, age, and aspirations. Others concern properties of the choice environment, such as its complexity (number of options), variance of outcomes, domain (gain vs. losses), and competition. One particularly interesting property of the choice environment is the order of decision problems within a study. Do individuals adapt their sample size across a sequence of problems because they build up expectations about how the problems are structured (e.g., the number of outcomes per option) and harness these priors to curtail search? One analysis has indeed observed such an effect (Lejarraga et al., 2012).

Taking advantage of our extensive database, we reexamined, where possible, the suggested moderators listed in Table 4 and also

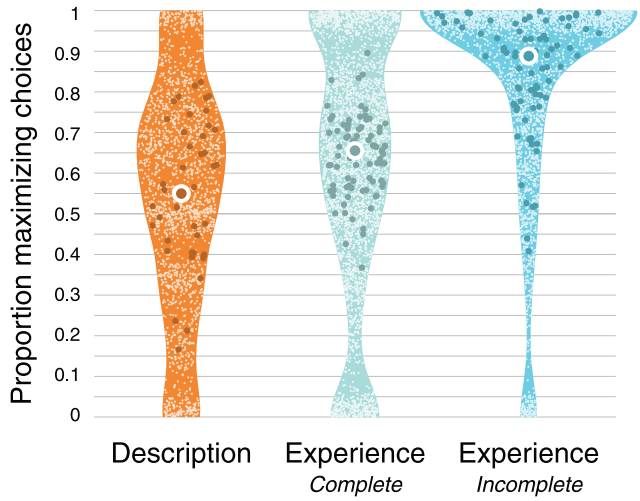

Figure 6. A description-experience gap in value maximization. Violin plots of the proportion of maximizing choices (of either expected value or experienced mean returns) for decisions from description and decisions from experience, separately for cases in which all possible outcomes were (not) sampled: "complete" versus "incomplete" experience. Every bright dot represents the result for one participant; every dark dot, for one data set. Circles represent the median result across data sets. See the online article for the color version of this figure.

Table 4
Determinants of Sample Size Identified to Date

Moderator	Article	Manipulation	Sample size[a] Treatment	Sample size[a] Control	Absolute difference
External					
Competition	Phillips et al. (2014)	Social competition (yes vs. no)	1	18	17
Complexity	Hills et al. (2013)	32 vs. 2 options	34	5	29
	Noguchi and Hills (2016)	32 vs. 2 options	51/38	6/4	
	Frey et al. (2015)[b]	8 vs. 2 options	113	41	72
Domain	Lejarraga et al. (2012)[c]	Loss vs. gain	11	9	2
Incentives	Hau et al. (2008)	Incentives × 10	33	11	22
Problem order	Lejarraga et al. (2012)	1st vs. 30th problem	25.5	9.1	16.4
Social context	Fleischhut et al. (2014)	Ultimatum game vs. standard paradigm	8	24	16
Variance	Lejarraga et al. (2012)[c]	Variance experienced (yes vs. no)	16	11	5
	Mehlhorn et al. (2014)	Variance experienced (yes vs. no)	15	3	12
Internal					
Affect	Frey et al. (2014)	Fearful vs. happy (study 1/study 2)	45/45	28/6	17/39
Age	Frey et al. (2015)[b]	Younger vs. older adults	46	58	12
Aspirations	Wulff et al. (2015b)	Long- vs. short-run	34	23	11
Numeracy	Lejarraga (2010)	High vs. low	23	15	8
Rational ability	Lejarraga (2010)	High vs. low	22	18	4

		Variable	Correlation[d]
Fluid intelligence	Frey et al. (2015)	DSST[e] & 2 options	<.1
	Frey et al. (2015)	DSST & 8 options	~.2 to ~.4
Working memory	Rakow et al. (2008)	Digit span	.38
	Wulff et al. (2015a)	Operation span[f]	.04
	Wulff et al. (2015b)	Operation span	−.19/.13/.19[g]

[a] Central tendency of sample size as reported in the article (numbers are rounded). [b] Results from this study were derived from the Bayesian regression results. [c] Lejarraga, Hertwig, and Gonzalez (2012). [d] Correlation between sample size and fluid intelligence and working memory, respectively. [e] Digit–symbol substitution task; see Wechsler (1981). [f] See Unsworth, Heitz, Schrock, and Engle (2005). [g] Article reports hierarchical estimates. None of the correlations were significantly different from 0.

examined some new ones. Specifically, we analyzed the moderating effect of (a) the structure of the decision problem (i.e., choice between a risky and a certain option vs. two risky options); (b) the probability of the rarest event ($z(p_{min})$); (c) the domain of the decision problem (gain vs. loss vs. mixed); (d) the order of decision problems (whether a problem was in the first or second half of the sequence); (e) the relative size of the expected values (calculated as $[z(ev_A) - z(ev_B)]$); (f) the absolute size of the expected values $[|z(ev_A) + z(ev_B)|]$; (g) the relative coefficient of variation $[z(cov_A) - z(cov_B)]$; and finally, (h) the absolute coefficient of variation $[|z(cov_A) + z(cov_B)|]$. The z-standardization was carried out on the level of the individual study. In light of the substantial role of problem structure, we present the results for risky versus safe and risky versus risky choices separately.

Table 5 reports the results of the moderator analysis for sample size. We found two robust moderators (insofar as the results for both structures of the decision problems converge). First, individuals sample more when the problem involves losses. Second, they sample less when the problem is in the second half of the sequence than they do when it is in the first half. Given that order had no systematic effect on the size of the gap, the latter finding may likely be attributable to higher efficiency rather than to fatigue. We also found three moderators specific to the problem structure (for details, see Table 5). We found no evidence that sample size is affected by the variance of the problems. These results suggest two conclusions. First, the characteristics of the problem—in particular, problem structure—impact search as well as choice. Second, and more generally, individuals adapt their sample size flexibly, possibly based on expectations they form with growing experience.

Table 5
Moderator Analysis of Sample Size

Moderator	Effect on samples per unit or standard deviation[a] Risky vs. safe	Risky vs. risky
Rarity[b]	−.56, $p < .076$	**−1.35, $p < .001$**
Domain = gain[b]	**−1.58, $p < .001$**	**−1.82, $p < .001$**
Problem order[b]	**−2.36, $p < .001$**	**−1.86, $p < .001$**
Relative EV[c]	.47, $p = .138$	**−.68, $p < .001$**
Absolute EV	**1.02, $p < .001$**	**−.28, $p = .02$**
Relative COV[d]	.18, $p = .466$	.00, $p = .983$
Absolute COV	−.21, $p = .349$	−.01, $p = .913$
Median sample size	14	22

Note. We also included the incentivization of the choice in the analysis, but No effects reached significance. For the sake of brevity, we omitted this variable from the analysis.
[a] We report effects per standard deviation for continuous variables and effects per unit change for binary and binarized variables. [b] Binary variable. [c] Expected value. [d] Coefficient of variation (i.e., standard deviation divided by expected value).

Is the Description-Experience Gap Solely Attributable to Sampling Error?

People's sample sizes are relatively small (see Figure 4), causing systematically distorted experienced-based representations of the true probabilities (see Figure 5). But is this restrained search alone responsible for the description-experience gap? To examine this possibility, experimenters have employed several methods to render description- and experienced-based information more similar or even identical, including raising the stakes, yoking, and fixing sampling sizes. Table 6 lists the methods applied and the results obtained. Two key results are noteworthy. First, the large majority of studies found a gap that was consistent with the discrete underweighting of rare events. Second, results were very heterogeneous, with gap sizes ranging from a minimum of 3.5 percentage points (Camilleri & Newell, 2011b) to a maximum of 26.1 percentage points (Camilleri & Newell, 2011b), and a weighted average of 14.8 percentage points as determined by a random effects model (Borgenstein et al., 2009).

One likely reason for this heterogeneity is that some of the methods came with unanticipated "side effects." For instance, translating one person's experienced outcome distributions into another person's described outcome distributions can result in trivial choices, such as the choice between "0 with certainty versus 3 with certainty;" in this case, the sampler had not experienced the rare event 32, and it is obvious that under these circumstances experience and yoked description elicit the same choice (see Hau et al., 2010). Another "side effect" is that requiring people to sample a specific N (e.g., 100 draws) takes the decision of when to terminate sampling away from the decision maker. As we show later, this affects the nature of the sampling process—for instance, by rendering optional stopping impossible (see Hogarth & Einhorn, 1992; Wulff & Pachur, 2016). Finally, various methods (i.e., pseudorandom sampling and sampling without replacement) inevitably introduce autocorrelations, and the experienced distributions thus cease to be iid random variables (see Estes, 1959; Plonsky et al., 2015; Restle, 1970). To conclude, the various attempts to render description and experience more similar by either reducing or eliminating sampling error have been found to reduce but not eliminate the description-experience gap. Furthermore, some methods exacted "side effects," rendering the results difficult to interpret.

Our meta-analytical approach enables us to examine the impact of equivalence in described probabilities and experienced relative frequencies without using the aforementioned methods, thus avoiding their difficulties. Taking advantage of our large pool of trials (N = 41,223; autonomous and matched sampling), we identified a set of trials in which individuals' experienced relative frequencies of the nonzero outcome were identical with or approximated the true probabilities (specifically, within a margin of error of 10% of the true probabilities; meaning that experienced frequencies of 9% and 11% would be treated as identical with the true probability of 10%). Following Camilleri and Newell (2011b), we were thus able to study the description-experience gap as a function of having equivalent information about the probability of outcomes. Camilleri and Newell defined description-experience equivalence with regard to the relative experienced frequency of the rare event (rather than the nonzero outcome). Our definition has the advantage of keeping the range of the experienced average value of an option constant across problems (i.e., 10% above and 10% below the expected value of the option).

Figures 7A and 7B illustrate the findings. For two frequently used decision problems, they plot the proportion of choices in favor of the risky option, separately for experience and description and for levels of information equivalence. If sampling error, that is, experienced frequencies deviating from true probabilities, were the

Table 6
Methods Used to Reduce or Eliminate Sampling Error in Studies on Decisions from Experience and the Description-Experience Gap

Method	Gap[a]	Sampling error
Higher monetary incentives		
Hau et al. (2008), Exp. 2	13%	Sampling error reduced
Fixed large sample sizes (e.g., N = 100)		
Camilleri and Newell (2011a)	14.8%	Sampling error reduced
Hau et al. (2008), Exp. 3	14.8%	
Hau et al. (2010), Exp. 1	19.8%	
Fixed large sample sizes and sampling without replacement		
Ungemach et al. (2009), Exp. 1	18.7%	Sampling error eliminated
Ungemach et al. (2009), Exp. 2	22.7%	
Pseudorandom sampling algorithm		
Camilleri and Newell (2011b), Exp. 1	26.1%	Sampling error reduced or eliminated
Camilleri and Newell (2011b), Exp. 2	4.1%	
Yoking of decisions from description to experience		
Rakow et al. (2008)	4.8%	Sampling error eliminated
Hau et al. (2010)	10%	
Trials of equivalent experience		
Camilleri and Newell (2011b), Exp. 1	3.5%	Sampling error largely eliminated
Weighted mean[b]	14.8%	

Note. Another method, devised by Hadar and Fox (2009), presents participants with information about all possible outcomes after sampling. Because the method has not yet been used in the context of comparable decisions from description, we excluded it from this analysis.
[a] Results reported in the articles. [b] Based on a meta-analytical random effects model (see Borgenstein et al., 2009).

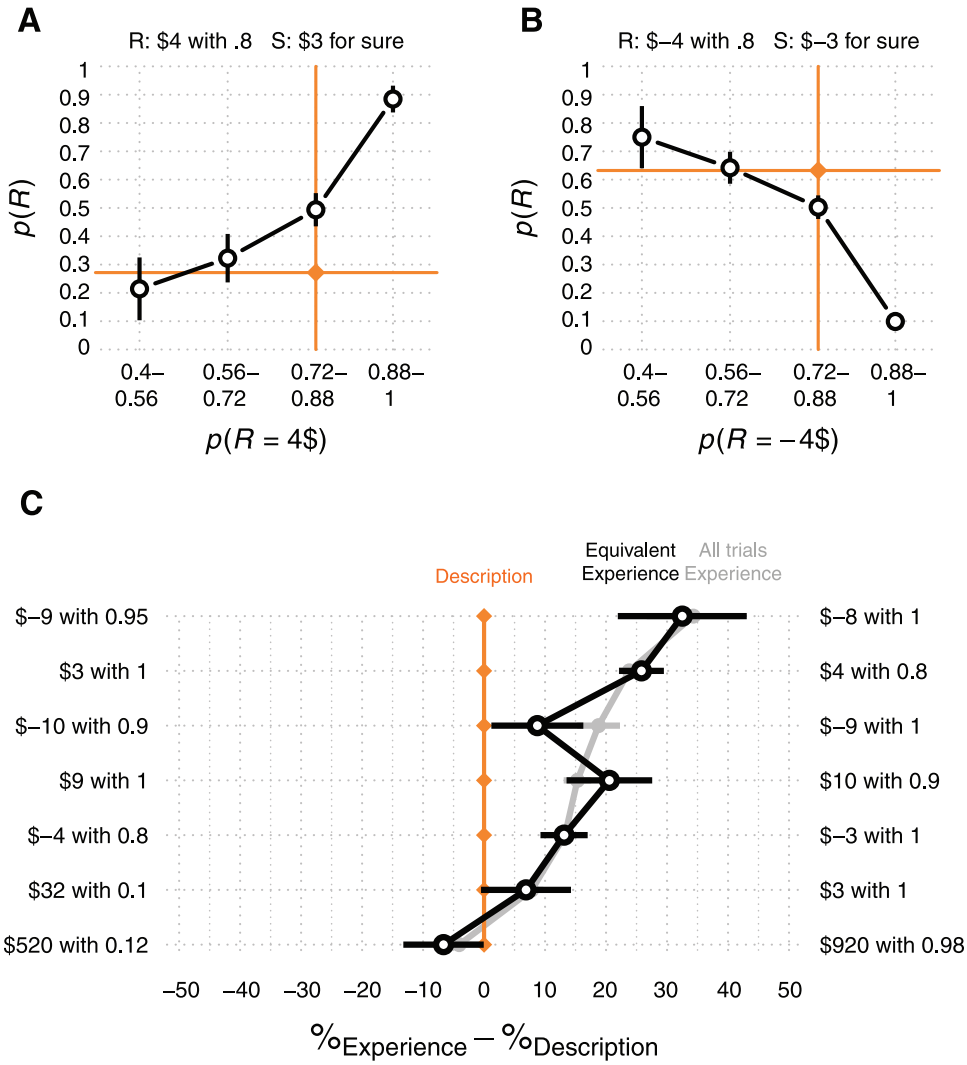

Figure 7. The description-experience gap as a function of the discrepancy between experienced frequencies and true probabilities. Panels A and B plot the proportions of risky choices p(R) against the relative frequency of the nonzero event, for example, p(R = $4) for a frequently used problem and its reflection. Panel C shows the results for all problems for which a defined number of equivalent trials could be obtained (see text). The orange diamonds (and lines) indicate choice proportions for decisions from description. The black circles indicate decisions from experience as a function of equivalence. Gray lines indicate the choice proportions aggregated across all experience-based trials. Error bars ± standard error of the gap. See the online article for the color version of this figure.

sole cause of the description-experience gap, then it should disappear once they are equivalent. As Figures 7A and 7B show, this not the case. There is a gap even when the experienced frequencies closely track the true probabilities. Moreover, as Figure 6 shows, the magnitude of the gap does not necessarily decrease with decreasing sampling error.

Figure 7C plots the results for six decision problems in which we found a "sufficient" number of trials to conduct the equivalence analysis. Across autonomous and matched sampling data sets (with matching description data), we identified a total of 3,681 (8.9%) trials. They were unevenly distributed across problems and data sets. In order to be able to draw reliable inferences, we therefore focused the analysis on problems for which at least 40 trials were identified, resulting in a total of 690 (1.7%) trials. As Figure 7C shows, in 5 out of 6 problems, a description-experience gap (consistent with the discrete underweighting of rare events) emerged, even when description and experience were equivalent (as defined). This finding was corroborated using a linear-mixed effects analysis of all 3,681 trials that controlled for study and participant random effects (discrete underweighting: $z = 10.34$, $p < 001$; CPT-based operationalization: $z = -11.74$, $p < .001$). Another interesting result is that even in those problems where experience closely tracks stated probabilities (and choice difficulty is thus comparable), there is a moderate gap in the maximization rate (68.9 vs. 63.0%).

To conclude, any experience will necessarily represent "just" a sample from the outcome distribution. Therefore, experience enters sampling error into choice. Is sampling error the sole mechanism behind the description-experience gap in the sampling paradigm? Our results suggest two conclusions: First, sampling error is indeed a key mechanism. Across studies, we found that people's sample size was relatively small, causing them to overlook events in about one third of all trials; the overlooked events were often the rare and consequential ones. Second, reducing or eliminating sampling error does not automatically reduce or eliminate the description-experience gap. Across the various methods used to approximate equivalence between stated probabilities and experienced frequency, there was still a substantial description-experience gap (see Table 6). Furthermore, a substantial gap still emerged in trials with largely equivalent information (see Figure 7). The latter result suggests that sampling error may be *sufficient* but not *necessary* for a gap to emerge. In other words, there are likely to be other mechanisms involved. We now turn to another candidate mechanism: recency.

Does Recency Contribute to the Description-Experience Gap?

A stream of experience needs to be integrated into some kind of mental representation of the available options. In decisions from experience—in contrast to decisions from description—it is up to the decision maker to perform such mental integration and abstraction. There are a number of reasons why each experienced episode may not be weighted equally in this process. For instance, the environment may be nonstationary and people may therefore discount past episodes more than recent episodes (e.g., Plonsky et al., 2015; Speekenbrink & Shanks, 2010). In stationary environments, people may weight past episodes less than recent episodes because of memory decay (e.g., Brown, Neath, & Chater, 2007; Murdock, 1962). Regardless of its causes, the attribution of less weight to past episodes than to recent episodes may contribute to a discrepancy between description and experience. In what follows, we examine the potential role of recency.

Within research on choice and judgment, the recency effect occurs when more recent outcomes in a stream of episodes have more predictive power for the individual's ultimate choice (judgment) than earlier outcomes do (e.g., Hogarth & Einhorn, 1992). One way to think about the impact of recency is as a process that further limits what is already a relatively small sampling size in decisions from experience (see Figure 4). In the extreme case, truncating a stream of experience to the most recent episodes will increase the likelihood of overlooking or underexperiencing rare events. In the less extreme case, cropping a stream of experience will reduce the impact of earlier episodes relative to that of recent episodes. In both cases, recency can bring about a lack of equivalence in experienced and described information.

Recency has long been entertained as a possible contributor to the description-experience gap in the sampling paradigm (Hertwig et al., 2004) and the partial-feedback paradigm (Barron & Erev, 2003). However, some researchers have expressed doubts about its existence and importance (Abdellaoui et al., 2011). Findings on its effect in the sampling paradigm are mixed (e.g., Hau et al., 2008; Hertwig et al., 2004; Rakow et al., 2008; Rakow & Rahim, 2010; Ungemach et al., 2009): Out of 21 studies, 10 found recency, nine found no recency, and two found primacy, that is, the opposite of recency (see Table 7). However, these results offer some indication that the existence of a recency effect depends on how autonomous sampling is. Specifically, recency appears more likely to occur under autonomous sampling (7/9) than under regulated sampling (3/8). Using these findings as a starting point, we next examined whether this regularity is robust and studied its implications.

How Robust is Recency and When Does it Occur?

Does the existence of a recency effect hinge on whether it is the individual or the experimenter who decides when to terminate sampling? To find out, we analyzed the occurrence of recency across the three variants of the sampling paradigm: *autonomous*, *matched*, and *regulated*. Furthermore, we employed three different measures of recency (measured on the level of individual search trials), all of them implementing variants of the following four steps: (a) Divide the stream of experience for each of the payoff distributions into a primacy and a recency set. (b) For the primacy and recency sets separately, determine which of the payoff distributions has the higher average outcome. (c) Calculate how often the payoff distribution with the higher average mean, as measured in the primacy sets and recency sets, respectively, has been chosen. (d) Evaluate whether the final choice is better explained in terms of an individual's experience in the early or late part of the stream of outcomes. The variants differ in how step (a) is implemented: The *recency-within-option* variant divides the stream of experience in two halves (primacy vs. recency sets) separately for each payoff distribution. The *recency-across-options* variant divides the entire stream of experience (as it occurred) in two halves. Finally, the *mirror-image* variant assigns all experiences sampled from the two payoff distributions before the second switch (that is, all experiences gathered before an individual returns for the first time to a distribution she has already sampled from) to the primacy set, and, following the same logic, all experiences gathered after the second but last switch to the recency set.

Figure 8 plots the observed predictive power of recency (i.e., how often the higher mean in the recency set predicts the final choice) relative to that of primacy. In the autonomous (upper panel) and matched data sets (autonomous with pseudorandom sampling; middle panel), a clear recency effect emerged across all three measures of recency. In the regulated data sets (lower panel), however, no recency effect occurred. These results suggest that the existence of recency is a function of the sampling regime. If the sampler's autonomy is curtailed and he or she is forced to sample up to a certain N (as is the case in the regulated data sets), no recency is likely to occur. Before we turn to possible reasons for this finding, let us first consider two objections to our analysis.

First, because we split the sequence in two parts and pitted the possible effects of recency and primacy against each other, it is conceivable that the null effect for the regulated sampling data sets may result from equally strong primacy and recency effects; thus, neither effect can emerge as the winner in a direct comparison. We therefore conducted a more fine-grained anal-

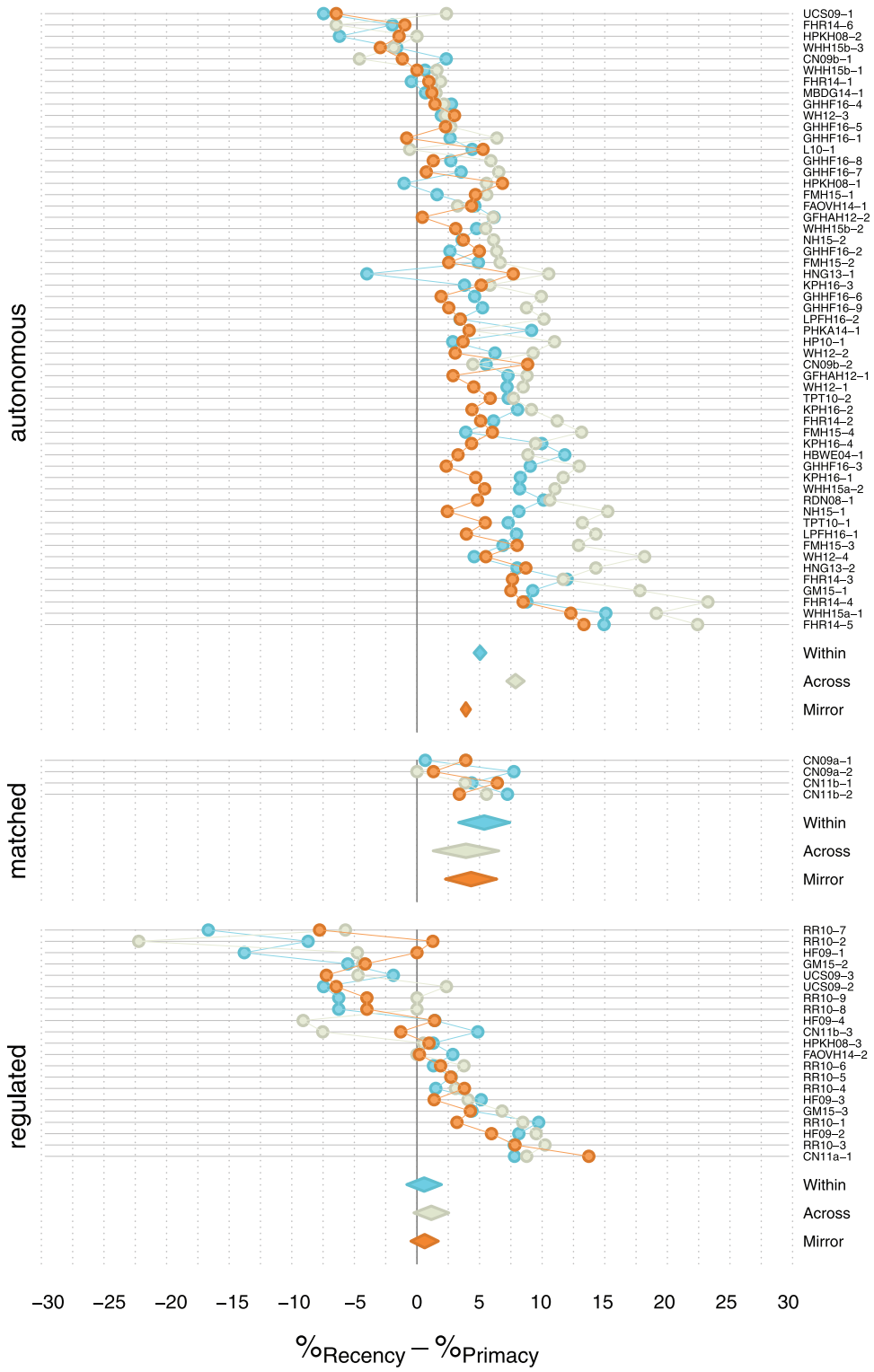

Figure 8. Analysis of recency in the sampling paradigm. Blue, beige, and orange points represent the results obtained for a given data set using the within-option, across-option, and mirror-image method, respectively. Diamonds and their widths represent the estimates and standard errors from a random effects meta-analysis. See Table A1 in Appendix A for a key to the study abbreviations. See the online article for the color version of this figure.

Table 7
Summary of Previous Findings on Order Effects (See Appendix D for Study-by-Study Details; Results are Reported as Interpreted by the Authors)

	Paradigm type (*N* data sets)			
	Autonomous (9)	Matched (4)	Regulated (8)	Total %
Recency	7	—	3	47
No effect	2	4	3	43
Primacy	—	—	2	9

ysis distinguishing between three parts of the sequence (see Appendix E) and found two results. First, there was no evidence for any order effects in the regulated sampling data sets. Second, there was clear evidence for a recency effect but no primacy effect in the autonomous and matched sampling data sets.

Second, the sample sizes in the regulated data sets in many cases exceeded the median of 20 samples in the autonomous data sets (see Figure 4). As a consequence, both primacy and recency sets in the regulated data sets are more likely to veridically reflect the underlying payoff distributions than in the autonomous data sets, thus limiting in theory the number of cases in which the primacy and the recency effect could result in distinct predictions. Table 8, however, shows that, relative to autonomous and matched sampling data sets, regulated sampling produced equally many—if not more—trials in which recency could have been detected ("% discriminatory"). In other words, the difference in the occurrence of recency does not stem from systematically different sampling sizes across the variants of the sampling paradigm.

Why Does Recency Depend on the Sampler's Choice to Terminate Search?

Two explanations of recency have been proposed in research on decisions from experience: *value updating* and *memory limitations*. Drawing on the notion of reinforcement learning (Hilgard & Marquis, 1940; Sutton & Barto, 1998), the value updating explanation assumes that the episodes (draws) in the stream of experience are continuously integrated into composite values of the options (Barron & Erev, 2003; Bush & Mosteller, 1955; Estes, 1959; Frey et al., 2015; Hertwig et al., 2006; Hogarth & Einhorn, 1992; March, 1996; Rescorla & Wagner, 1972). Such an ongoing learning process can produce recency or primacy, depending on how the updating unfolds (Hertwig et al., 2006). However, making the minimal assumption that each episode, regardless of its position in the sequence, receives a constant weight, will always produce recency (Anderson & Hovland, 1957; Hogarth & Einhorn, 1992).

Memory-based explanations of recency in decisions from experience assume that the stream of experience or some proxy will be explicitly stored in memory (Ashby & Rakow, 2014; Gonzalez & Dutt, 2011; Hawkins, Camilleri, Heathcote, Newell, & Brown, 2014; Lejarraga, Dutt, & Gonzalez, 2012; Lin, Donkin, & Newell, 2015). A representation of the stream of experience can, however, still give rise to strong recency (Atkinson & Shiffrin, 1968; Murdock, 1962), caused by limited capacity of working memory (Baddeley, 1986; Cowan, 2001; Miller, 1956), decay of the memory trace (e.g., Anderson & Lebiere, 1998), or interference (e.g., Farrell & Lewandowsky, 2002; for a review, see Oberauer, Farrell, Jarrold, & Lewandowsky, 2016).

Can value updating and memory limitations explain the occurrence of recency in the autonomous (and matched) data sets and the lack of recency in the regulated data sets? Because sample sizes are, on average, larger in regulated than in autonomous sampling, memory limitations would predict, in contrast with our findings, *more* recency in regulated data sets. Value updating predicts the same amount of recency for autonomous and regulated sampling and thus also fails to provide a good account of our

Table 8
Analysis of Recency in the Sampling Paradigm

Variant	Method	Aggregate effect $\%_{Rec}-\%_{Prim}$[a]	Mixed effects analysis (trial level)	% valid[b]	Median N_{Rec}/N_{Prim}[c]	% discriminatory[d]
Autonomous	Within	5.46	$z = 17.08, p < .001$	95	10/10	30
	Across	7.55	$z = 19.95, p < .001$	49	10/10	17
	Mirror	4.20	$z = 13.43, p < .001$	95	14/11	16
Matched	Within	5.29	$z = 2.72, p = .007$	96	7/6	28
	Across	4.09	$z = 1.90, p = .058$	56	6/6	18
	Mirror	4.12	$z = 2.15, p = .032$	95	4/6	18
Regulated	Within	.01	$z = .72, p = .474$	89	10/10	40
	Across	.00	$z = .41, p = .685$	56	10/10	26
	Mirror	.01	$z = 1.02, p = .307$	90	6/18	22

[a] Difference in proportion of choices consistent with the better payoff distribution in the recency/primacy set. [b] Percentage of times in which both primacy and recency sets rendered predictions. [c] Median number of samples in the recency/primacy sets. [d] Percentage of times in which primacy and recency sets predicted the choice of different payoff distributions.

findings. How, then, can the recency effect's contingency on the sampling instruction be explained?

Let us consider two other possible explanations. First, autonomous and regulated sampling may trigger different valuation mechanisms. In their influential article on order effects in belief updating, Hogarth and Einhorn (1992) found that recency occurs only in "step-by-step tasks." These tasks require decision makers to recurrently evaluate an option after each episode in a stream of experience (i.e., piecemeal valuation). By contrast, "end-of-sequence tasks" leave it to decision makers to evaluate the option either in a piecemeal way or once at the end of the experience. End-of-sequence tasks predominantly produced primacy. To explain this difference, Hogarth and Einhorn (1992) proposed that step-by-step tasks and piecemeal valuation cause a value-updating process (conducive to recency). End-of-sequence tasks, in contrast, permit a range of processes, including the explicit storage of all episodes (Lindskog, Winman, & Juslin, 2013). Mapping the distinction between end-of-sequence and step-by-step tasks onto autonomous and regulated sampling, one may hypothesize that leaving the decision of when to terminate search to the sampler causes piecemeal valuation and, by extension, recency. In contrast, revoking the autonomy to terminate search in regulated sampling turns this paradigm into an end-of-sequence task that permits other processes and, by extension, primacy, recency, or even no effects (see Table 1 in Hogarth & Einhorn, 1992; see Wulff & Pachur, 2016).

A second possible explanation for recency's contingency on the sampling instruction is *optional stopping* (Fried & Peterson, 1969; Wald, 1947). In optional stopping, the stopping decision can be made at any point in the sampling process, and the individual uses properties observed in the sampled data to determine when to stop sampling (e.g., the occurrence of a rare, consequential event or arriving at a decision threshold; see Hertwig & Pleskac, 2010; Markant, Pleskac, Diederich, Pachur, & Hertwig, 2015). In fixed stopping, in contrast, the sample size is specified in advance. In random stopping, search can be stopped at any arbitrary point, with this termination decision being probabilistic and independent of the data at hand. Optional stopping requires that the choice about when to terminate search resides with the decision maker. This is the case in autonomous sampling. In regulated sampling, in contrast, this choice lies with the experimenter. The distinction between optional, fixed, and random stopping is important because only the first can produce recency. In Appendix F, we demonstrate that optional stopping can indeed lead to recency by implementing two possible optional stopping strategies.

More experimental work is needed to test which of these two explanations—optional stopping versus piecemeal and end-of-sequence valuations—best explains the results shown in Figure 8. For instance, one experimental approach barely implemented thus far is to determine the quality of people's knowledge about experienced outcomes and their relative frequencies. A piecemeal evaluation (as may occur in the step-by-step task of uninstructed sampling; Hogarth & Einhorn, 1992) implies that no explicit representation of the sequence of experience exists. Consequently, one would expect participants to have difficulties answering questions about experienced outcomes and relative frequencies. However, the limited evidence available suggests that people, at least when their opinions are aggregated, have a reasonable sense of the relative frequencies with which they experienced outcomes (Bradbury, Hens, & Zeisberger, 2014; Camilleri & Newell, 2009a; Fox & Hadar, 2006; Hau et al., 2008; Kaufman, Weber, & Haisley, 2012; Lejarraga, 2010; Ungemach et al., 2009), consistent with the notion of automatic encoding of frequency information (see Hasher & Zacks, 1984; Zacks & Hasher, 2002). Optional stopping, because it makes no assumption about how experience is represented in memory, would at least not be inconsistent with these initial results.

To further explore the possibility of optional stopping, we also analyzed the equivalent of a *gaze-cascade effect* in decisions from experience. When observers are asked to choose between two stimuli (e.g., which of two faces is more attractive), gaze is at first equally distributed between the options before, 1 or 2 s before choice, progressively shifting toward the ultimately chosen option (Shimojo, Simion, Shimojo, & Scheier, 2003). The equivalent of this gaze-cascade effect in decisions from experience would be an increase in the likelihood of sampling from the ultimately chosen option shortly before choosing. Researchers have sought to explain the gaze-cascade effect in terms of optional stopping mechanisms (for an overview, see Mullett & Stewart, 2016). A cascade-like effect in decisions from experience may thus also suggest the involvement of optional stopping mechanisms. To test this possibility, we calculated separately for autonomous and regulated sampling the relative frequency with which individuals sampled from the chosen option as a function of the draw's relative position in the sequence. In order to be able to detect a progressive shift toward the chosen option, one requires sequences that are not overly short. For this reason, we included only sequences of at least 10 samples (which also permitted us to analyze the progression of sampling behavior for 10 different bins in the sequence).

Figure 9 plots the likelihood of sampling from the ultimately chosen option for autonomous and regulated sampling. We found two results: First, there was a cascade-like effect in autonomous sampling, starting between the eighth and ninth bin. Second, in regulated sampling, there was a sampling bias toward the ultimately chosen option starting around the fifth bin. This difference could be interpreted as consistent with optional stopping in autonomous sampling but not regulated sampling.

To conclude, in the sampling paradigm, people tend to rely on small samples, and thus risk missing events that are rare but consequential (Pleskac & Hertwig, 2014; Taleb, 2010). This risk would be even greater if order effects such as recency were to cut the sequence of the "operative" experience even shorter. After the original observations of recency by Barron and Erev (2003) and Hertwig et al. (2004), subsequent findings were mixed. Our analysis clarifies why results have been so heterogeneous. The recency effect (i.e., when the more recent segment of experience predicts final choice better than the initial segment does) occurs only when the sampler is given autonomy over sample size and stopping. When the experimenter regulates the sampling size, no recency occurs (see Figure 8). We discussed two possible explanations, neither of which is fully satisfying at this point. Finally, consistent with optional stopping, we observed a gaze-cascade-like effect in autonomous sampling.

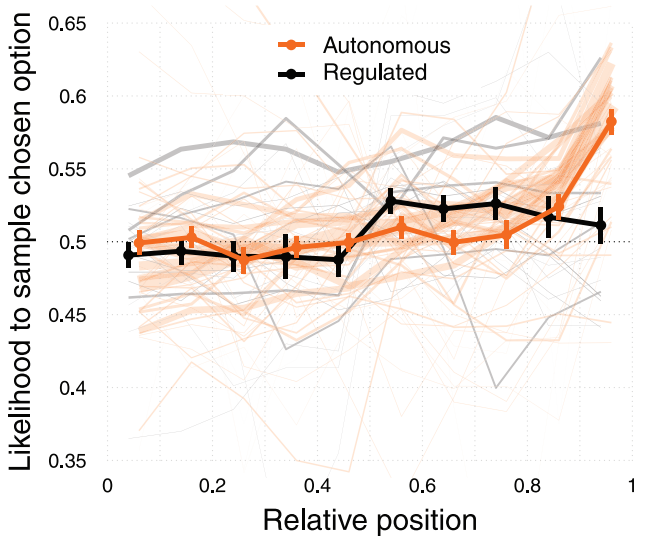

Figure 9. The gaze-cascade-like effect in decisions from experience. The figure shows, separately for autonomous and regulated sampling, the likelihood of sampling from the (ultimately) chosen option as a function of the draw's relative position in the sampling sequence (for all sequences of length ≥10; 1,024,307 autonomous and 158,883 regulated sampling trials). The lines in the background show the results separately for every autonomous and regulated dataset. The error bars represent the standard error of the mean. See the online article for the color version of this figure.

Does Reversed Probability Weighting "Explain" the Description-Experience Gap?

Prospect theory, the most influential descriptive theory of risky choice, was originally developed for "simple prospects with monetary outcomes and *stated probabilities*" (Kahneman & Tversky, 1979, p. 274; emphasis added)—in other words, for decisions from description. The theory makes an important assumption about the impact of stated probabilities, namely, that outcomes are not linearly weighted by their probabilities but that "decision weights are generally lower than the corresponding probabilities, except in the range of low probabilities" (p. 263). Specifically, low-probability events tend to be overweighted (originally discussed by Edwards, 1954, see his Figure 3). The notion of overweighting of low probabilities (i.e., rare events) has been crucial as it could, for instance, help to explain a long-standing puzzle of the utility function, namely, that people purchase both insurance policies (risk aversion) and lottery tickets (risk seeking; Friedman & Savage, 1948; Markowitz, 1952). In both cases, people seem to overweight rare events—in one case, a rare large loss; in the other, a rare large gain.

With the enormous impact of prospect theory (Kahneman & Tversky, 1979), an inverse S-shaped probability weighting function has widely been taken for granted as a psychological regularity (we return to this point shortly). It is perhaps against this background that the possibility of reversed probability weighting in experienced-based choice has captured researchers' interest. Barron and Erev (2003) and Hertwig et al. (2004) suggested the following conclusion about the description-experience gap:

> Differences in choices were consistent with the assumption that in decisions from experience, rare events had less impact than they deserved on the basis of objective probability (and in decisions from description, rare events had more impact than they deserved; Hertwig et al., 2004, p. 535; see also Weber et al., 2004).

It is important to note that Hertwig et al. (2004) and Barron and Erev (2003)—like Kahneman and Tversky (1979)—inferred the potential probability weighting from people's choices. That is, their conclusions were not derived from estimating a weighting function conditioned on people's actual experience (which was done later in the context of cumulative prospect theory; Tversky & Kahneman, 1992). In Hertwig et al. (2004), for instance, lotteries were selected such that if rare events were accorded less weight in experience than in description, systematically different patterns of choices would result. This approach had two consequences for the inferred weighting pattern in decisions from experience. First, the weighting was meant in an *as-if* sense (i.e., people behaved as if rare events had less impact then they deserved) and, second, the as-if weights referred to the *objective* probabilities of the outcome distributions and not to the relative frequencies with which people actually experienced the events.

In the wake of the original articles documenting the description-experience gap, researchers began to quantitatively measure decision weights in experience and to explore whether they are qualitatively different from those measured in description. In so doing, they zoomed in on the following question: What decision weights are attached to the relative frequencies that people experienced (rather than to the objective probabilities), and are they any different from those attached to stated (described) probabilities? It is important to note that even if description-based and experience-based decision weights were similar or even identical, choices could still be systematically different. The reason for this is that—because of sampling error, recency, or other factors—the relative experienced frequencies or their perception can deviate systematically from the objective (stated) probabilities.

Table 9 summarizes all published studies of probability weighting in decisions from experience relative to decisions from description included in our analysis. It seems fair to say that the conclusions drawn are fairly mixed, with 12 analyses reporting underweighting, 14 reporting overweighting, one analysis reporting (nearly) linear weighting, three reporting smaller overweighting in experience than in description, and two inconclusive analyses. Why are results so mixed? One reason is that the researchers implemented the sampling paradigm in many different ways. For instance, Camilleri and Newell (2011a) and Ungemach et al. (2009, Exp. 2), whose results were inconclusive, used regulated sampling, whereas Frey et al. (2015) and Ungemach et al. (2009, Exp. 1), who observed underweighting, used autonomous sampling. Abdellaoui et al. (2011) provided participants with an exhaustive list of outcomes (including outcomes they had not experienced) immediately after sampling, thus examining a combination of description and experience. Moreover, researchers have used diverse approaches to determine the probability weighting parameters (e.g., measurement, best-fitting parameters). Given this methodological diversity, the heterogeneity of results and conclusions is not too surprising. Furthermore and importantly, a universal shape of the weighting function is not necessarily to be expected—it does not even exist in decisions from description. Commenting on the conflicting results, van de Kuilen and Wakker (2011) wrote: "Although we believe that inverse-S is

Table 9
Findings on Experience-Based Probability Weighting in Decisions from Experience

Analysis	Description parameter	Experience parameter	Comments	Inferred weighting of rare event
Sampling paradigm				
Hau		$\gamma = .99$	TK92; reanalysis of Hertwig et al. (2004)	Linear weighting
Ungemach et al. (2009)	—	$\gamma^+ > 1$[a] $\gamma^- > 1$	CP[b]-TK92; Experiment 1—autonomous sampling	Underweighting
	—	$\gamma^+ = [0, 2]$ $\gamma^- = [0, 2]$	CP-TK92; inconclusive results Experiment 1—regulated sampling	—
	—	$\gamma^+ > 1$[a] $\gamma^- > 1$	CP-TK92; Experiment 2—experienced probabilities	Underweighting
	—	$\gamma^+ > 1$[a] $\gamma^- > 1$	CP-TK92; Experiment 2—judged probabilities	Underweighting
Camilleri and Newell (2011b)	—	$\gamma = [0, 2]$	CP-TK92; inconclusive results	—
Camilleri and Newell (2013)	$\gamma = [0, 2]$	$\gamma > 1$	CP-TK92; single-play	Underweighting
	$\gamma = [0, 2]$	$\gamma > 1$	CP-TK92; multiplay	Underweighting
Frey et al. (2015)	—	$\gamma^+ = 1.3$ $(\delta^+ = 1)$ $\gamma^- = 1.35$ $(\delta^- = 1)$	P98; Experiment 2; sample of younger adults	Underweighting
	—	$\gamma^+ = 1.03$ $(\delta^+ = 1)$ $\gamma^- = 1.05$ $(\delta^- = 1)$	P98; Experiment 2; sample of older adults	Underweighting
Lejarraga et al. (2016)	$\gamma = .89$ $(\delta = .96)$	$\gamma = .81$ $(\delta = .87)$	P98; monetary problems	Overweighting
	$\gamma = .20$ $(\delta = 4.33)$	$\gamma = .53$ $(\delta = .3.82)$	P98; medical problems	Overweighting
Glöckner et al. (2016)	$\gamma = .73$ $(\delta = .55)$	$\gamma = .56$ $(\delta = .55)$	GE87; reanalysis of Glöckner et al. (2012)	Overweighting
	$\gamma = .73$ $(\delta = .32)$	$\gamma = .55$ $(\delta = .39)$	GE87; Experiment 1	Overweighting
	$\gamma = .96$ $(\delta = .70)$	$\gamma = .55$ $(\delta = .48)$	GE87; Experiment 2	Overweighting
	$\gamma = .65$ $(\delta = .80)$	$\gamma = .42$ $(\delta = .79)$	GE87; Experiment 3	Overweighting
	$\gamma = .59$ $(\delta = .96)$	$\gamma = .91$ $(\delta = 1.04)$	GE87; reanalysis of Erev et al. (2010)	Overweighting but less pronounced relative to description
Kellen et al. (2016)	$\gamma = .66$ $(\delta^+ = .81, \delta^- = 1.53)$	$\gamma = .53$ $(\delta^+ = .71, \delta^- = 1.66)$	GE87	Overweighting
Markant et al.[c] (2015)	—	$\gamma = 1.41$ $(\delta = 1)$	P98; reanalysis of Erev et al. (2010)	Underweighting
	—	$\gamma = 1.15$ $(\delta = 1.61)$	P98; reanalysis of Hau et al. (2010), Experiment 1	Underweighting
	—	$\gamma = .92$ $(\delta = 1.3)$	P98; reanalysis of Hau et al. (2010), Experiment 2	Overweighting
Variants of the sampling paradigm				
Abdellaoui et al.[d] (2011)	$\gamma^+ = .65$ $(\delta^+ = .70)$ $\gamma^- = .73$ $(\delta^- = .78)$	$\gamma^+ = .66$ $(\delta^+ = .59)$ $\gamma^- = .74$ $(\delta^- = .67)$	GE87	Overweighting but less pronounced relative to description
	—	—	Nonparametric estimation	Overweighting but less pronounced relative to description
Camilleri and Newell (2011b)	—	$\gamma > 1$	CP-TK92; partial-feedback paradigm	Underweighting
	—	$\gamma > 1$	CP-TK92; full-feedback paradigm	Underweighting
Jarvstad et al.[e] (2013)	—	—	Qualitative evaluation; experienced frequencies	Underweighting
	—	—	Qualitative evaluation; probability judgments	Overweighting
Zeigenfuse et al.[f] (2014)	—	$\gamma = .7$ $(\delta = .3)$	P98	Overweighting
Kemel and Travers[g] (2016)	—	$\gamma = .68$ $(\delta = .68)$	GE87; complete information, monetary outcomes	Overweighting

Table 9 (*continued*)

Analysis	Description parameter	Experience parameter	Comments	Inferred weighting of rare event
	—	$\gamma = .59$ ($\delta = .78$)	GE87; complete information, temporal outcomes	Overweighting
	—	$\gamma = .74$ ($\delta = .66$)	GE87; incomplete information, monetary outcomes	Overweighting
	—	$\gamma = .63$ ($\delta = .83$)	GE87; incomplete information, temporal outcomes	Overweighting

Note. GE87 = Two-parameter weighting function introduced by Goldstein and Einhorn (1987). TK92 = One-parameter weighting function introduced by Tversky and Kahneman (1992). P98 = Two-parameter weighting function introduced by Prelec (1998).
[a] The γ parameter of cumulative prospect theory's weighting function governs the shape of the function. There are different implementations of this function (see Comments). For our purpose, it suffices to know that γ values larger than 1 suggest underweighting and values smaller than 1 suggest overweighting. [b] Contour plots (CP) depict the proportion of choices correctly predicted by various parameter combinations. The value or values reported indicate the range that produced the best results. [c] This study estimated the parameters using a joint model of choice and sampling data. [d] This study relied on a variant of the sampling paradigm that describes all possible outcomes to the participant before choice. [e] This study relied on a perceptuo-motor task and a qualitative evaluation based on several different implementations of prospect theory. [f] This study relied on a speeded and perceptual variant of the sampling paradigm called the Flash Gambling Task. [g] This study required individuals to sample from a single option. We report the results from the mixed model estimation based on observed frequencies (see Table 11 in Kemel & Travers, 2016).

the prevailing phenomenon [in studies with stated probabilities], it is certainly not universal" (p. 594).

For several reasons, we limit our treatment of nonlinear probability weighting in description and experience to a primarily qualitative review. First, aggregating across such a methodologically diverse set of studies would risk producing average parameters of questionable value (see, e.g., the divergent results obtained by Glöckner et al., 2016, and Markant et al., 2015, for the same data as Erev et al., 2010). Second, Figure 10, which illustrates our exploration of the probability weighting patterns as a function of problem structure, strongly suggests that overweighting or underweighting appears to be a function of problem structure (Glöckner et al., 2016). This means that the true probability weighting patterns across type of problems and, by extension, studies (see Table 9) are indeed distinct and defy simple aggregation. Third, in order to reliably measure the weighting parameters in each of the studies in our database, we would need to rely on large problem sets in which the properties of the options (i.e., probabilities and outcomes) tend to be relatively independent of each other (Broomell & Bhatia, 2014; Glöckner & Pachur, 2012; Myung & Pitt, 2009; Wulff & Pachur, 2016; Wulff & van den Bos, 2017). Most data sets in our database do not meet these criteria. Finally, there is also a conceptual hurdle relating to the certainty effect in decisions from experience. Ideally, a better conceptual understanding of this issue is required before more weighting analyses are conducted (see Appendix G).

One of our findings does speak to the weighting issue. When we focused on trials in which people experienced the outcomes with roughly the same frequency as the stated probabilities, choices were still systematically different and the gaps were mostly consistent with either underweighting of rare events—or at least with less overweighting of rare events (including linear weighting) than in decisions from description (see Figure 7). This analysis is informative but certainly does not settle the issue. We agree with van de Kuilen and Wakker's (2011) sentiment that "[m]uch about weighting functions remains yet to be discovered" (p. 594). This is true for the dynamics of probability weighting in both description and experience.

Last but not least, let us emphasize that there is an alternative to the notion of nonlinear probability weighting in decisions from experience. Figure 6 shows that in decisions from experience a large proportion of choices are consistent with maximization of the experienced mean return. This, in turn, is consistent with either linear weighting of probabilities or, more radically, with no weighting of probabilities at all. One simple strategy that enables people to consistently select the option with the higher experienced return is the natural-mean heuristic (Hertwig & Pleskac, 2008, 2010; see also Sutton & Barto, 1998). This heuristic sums up all experienced returns for each payoff distribution, divides that sum by the number of returns per distribution, and then chooses the deck with the larger mean return. Such a strategy would not require any explicit representation of probabilities, let alone any probability weighting. The same holds for reinforcement models or value-updating models.

General Discussion

About a dozen years ago, three studies found evidence of a description-experience gap in risky choice (Barron & Erev, 2003; Hertwig et al., 2004; Weber et al., 2004), observing that distinct modes of learning and decision making appear to result in systematically different choices about the uncertainties of life. Since then, numerous studies have examined the reality and causes of this gap, and new insights have raised new questions. For this reason, it seems timely and important to use meta-analytical methods to review what has been learned to date and, hopefully, to bring some of the debated issues closer to resolution. To this end, we reanalyzed nearly 70,000 decisions made from description and experience in the sampling paradigm, the experiential paradigm most often used to date. We considered three possible determinants of the description-experience gap: reliance on small samples, recency, and reversal of probability weighting.

Major Findings

The description-experience gap is a robust regularity (see Figure 2). Across two definitions of the gap, we found that experience- and description-based choices differed by about 9.7 and 13.4 percentage points, depending on how the gap was measured. Second, the magnitude of the gap is moderated by various factors,

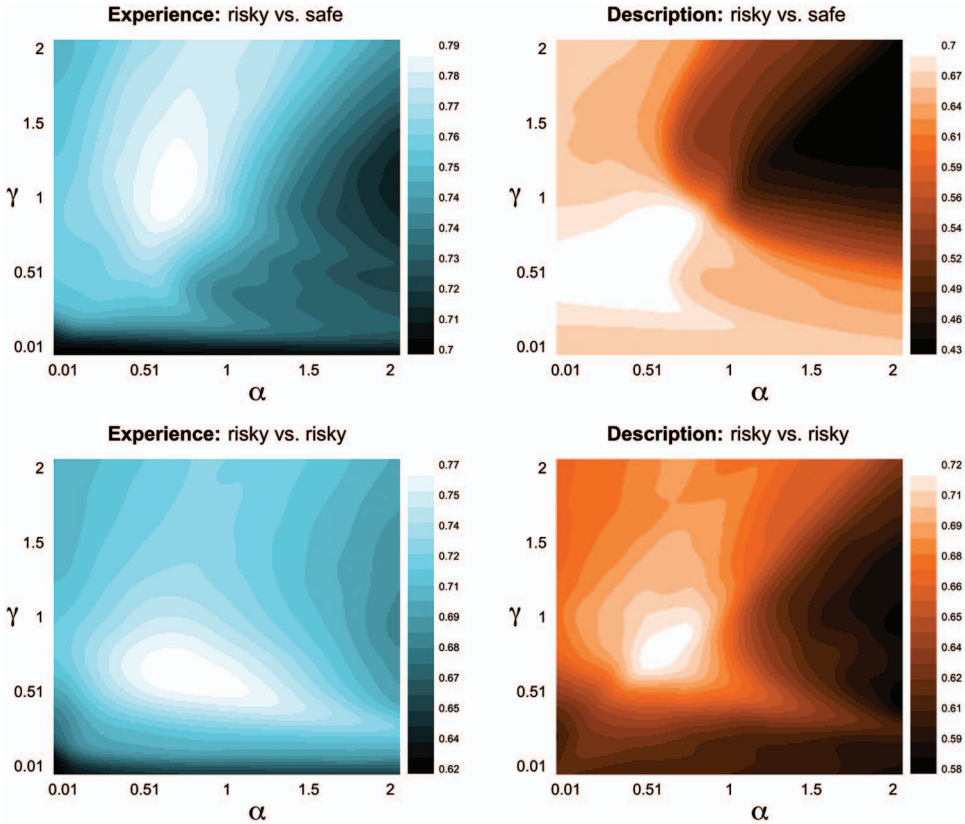

Figure 10. Contour plots showing the predictive accuracy of various parameter combinations of the two-parameter cumulative prospect theory model (α = [.01, 2], γ = [0.1, 2]; using the formulation of Tversky & Kahneman, 1992). The figure shows differences in probability (and outcome) weighting as a function of problem structure in experience and description. It also highlights regions of superior fit, suggesting linear weighting for risky versus safe choices in experience, and overweighting for the three other cases. However, for the reasons outlined in the text, these results should be treated with caution. In addition to the two-parameter variant of CPT, we implemented a more complex prospect theory model (including a loss aversion parameter) and found no improvement in the maximum predicted accuracy. Note that the results for decisions from experience were derived using the problems experienced. See the online article for the color version of this figure.

including the probability of the rare event, the problem's placing in the sequence and, most importantly, problem structure. Specifically, when a choice involved a risky and a safe option, the average magnitude of the gap was 20.2 percentage points; when two risky options were presented, it was about 7.5 percentage points (see Figure 3 and Table C3 in Appendix C). Third, people rely on small sample sizes—the median sample size was 14 across trials with a safe and a risky option and 22 across trials with two risky options. Small samples can systematically distort experience-based representations of options' objective properties (Figure 5, Table 5). In about one third of autonomous sampling trials (14,504 of the 40,246 trials), people did not experience at least one of the possible outcomes—typically the rare event). Sampling error does not fully explain the gap, however, as revealed by an analysis of those trials in which experienced frequencies and described probabilities converged (Table 6 and Figure 7). Fourth, we observed that experienced-based choices were substantially more consistent with the maximization of average returns than description-based choices were (see Figure 6). Fifth, we found that recency occurred only as long as the observer was able to terminate search of his or her own accord (see Figure 8). Without this autonomy, no recency was observed. Furthermore, we identified a gaze-cascade-like effect (see Figure 9), suggesting that recency stems from optional stopping (see also Appendix F). Finally, experienced relative frequencies and stated probabilities prompted different weighting functions for choices involving a risky and a safe option, but more similar functions for choices with two risky options (see Figure 10). This finding could be one key to explaining the large heterogeneity in published probability weighting analyses.

Determinants of the Description-Experience Gap

What do these findings mean for the three potential determinants of the description-experience gap in the sampling paradigm that have received most attention: reliance on small samples, recency, and reversed probability weighting? We discuss each potential determinant in turn. First, people indeed rely on samples that tend to systematically misrepresent the relative

frequencies of the experienced options (relative to the objective options). This misrepresentation can range from underrepresentation of the objective probability of rare events to unawareness of their existence. Relative to the distribution of true probabilities, the distribution of experienced relative frequencies is notably more polarized toward 1 and 0 (see Figure 5). This suggests that sampling error is an important contributor to the description-experience gap. However, the gap does not disappear when only those trials in which experience and description (closely) match are included in the analysis (see Figure 7). Sampling error is thus not the sole cause of the description-experience gap.

Past results concerning recency as a possible contributor to the description-experience gap have been mixed, with some researchers finding evidence of recency and others finding none. Our analysis revealed a likely reason for recency's ambiguous role in the description-experience gap: It seems to occur only under autonomous sampling and to be a consequence of people's strategic behavior (e.g., swift stopping after specific outcomes) rather than of memory decay. These findings raise an important issue beyond recency. Experience is, of course, not of a single kind. It can be the product of an active process in which the amount of information, the search process, its timing, and the search policy (Hills & Hertwig, 2010) are under the searcher's volitional control. For instance, a shopper in a department store self-directs the flow of experience, and the information that becomes available to her is the direct result of her actions, choices, and strategic behavior, which are, in turn, informed by expectations that have evolved across time. Alternatively, the amount, timing, and sequence of experience may be under the control of dispassionate nature (e.g., the experience of seasonality) or of other social agents (e.g., parents, teachers; see also Gureckis & Markant, 2012, on the distinction between active and passive information acquisition and learning). Our recency analysis suggests that it will be rewarding to identify and vary such attributes of experience (and description), and to examine which narrow or amplify the gap between the two forms of learning. In such future explorations, it will be worth keeping in mind that optional stopping alone cannot produce underweighting of rare events.

Our analysis also offers some structure to the mixed set of findings on probability weighting in decisions from experience (see Table 9). There are likely to be several reasons why past analyses resulted in such diverse weighting patterns, not least because of stark methodological differences in measurements. Distinguishing between problems with one safe and one risky option and problems with two risky options revealed that, in the former, experienced-based choice is best accommodated by linear to S-shape weighting functions. In the latter, experienced-based choice is best accommodated by an inverse S-shaped weighting function—the same function found in description-based choices (see Figure 10). This finding suggests that, in decisions from experience, perhaps more so than in decisions from description, the structure of the problem matters and may trigger different processes. This is also consistent with our finding of a large gap in problems with a risky versus safe option but only a small gap in problems with two risky options (see Glöckner et al., 2016). The former is the standard tool employed to measure people's risk preferences. Indeed, Weber et al.'s (2004) comprehensive search of the literature identified a total of 226 choice situations, each presenting a choice between a safe option with a certain outcome and a risky option with two potential outcomes.

Is the gap thus restricted to this one—albeit important—kind of problem structure? It is too early to tell. More studies are required in which the problem structure (e.g., number of options, number of events per option) is systematically manipulated. It may be that with more complex problems (e.g., multiple events or multiple options; Hills et al., 2013) requiring increasingly difficult-to-process descriptions, a gap will emerge—but possibly for reasons other than those identified here. Having said that, as long as rare events are involved in both of a decision problem's options (the gap has also been demonstrated for problem structures without rare events; see Ludvig & Spetch, 2011), it can be expected that choices between two risky options are less likely to produce a description-experience gap than choices between a risky and a safe option. In the former, both options may contain low-probability events to which all of the effects discussed—small samples, sampling errors, recency, and probability weighting—apply, rendering it possible that they cancel each other out. One way to gauge the diagnosticity of decision problems for the existence of the description-experience gap is to evaluate the overlap between discrete underweighting (of rare events) and the CPT-overweighting of rare events. If both predict the choice of the same option, then it is impossible to detect a description-experience gap. According to this criterion, 40% of the choices between two risky options in our database were nondiagnostic, relative to only 13% of the choices between a risky and a safe option. This can at least partly explain why the description-experience gap differs substantially across problem structures.

Let us briefly discuss one final issue. In problems with a risky and a safe option, the center of the density plot for experienced-based choice is located in the vicinity of linear weighting (see Figure 10). Furthermore, people are more likely to maximize the experienced mean reward than to maximize the expected value in description. Importantly, this also holds for cases in which description- and experience-based choices involve the same level of difficulty. Does this mean that experienced-based choice is more normative? This is an important issue and, in all likelihood, one with no straightforward answer. Fantino and Navarro (2012) reported that it is often the combination of description and experience that renders performance more optimal (see also Barron & Yechiam, 2009; Erev, Ert, Plonsky, Cohen, & Cohen, 2017; Jessup, Bishara, & Busemeyer, 2008). Schulze and Hertwig (2017) proposed the description-experience distinction as a possible response to the question of why grown-ups are often "so stupid about probabilities when even babies and chimps can be so smart" (Gopnik, 2014). In the experimental literature, babies have consistently emerged as good intuitive statisticians, capable of statistical learning and judgment, whereas adults' statistical judgments have often been found lacking (e.g., Kahneman, 2011). One possible key to this conundrum is that babies, unlike adults, cannot (yet) use symbolic descriptions of probabilistic information. Consequently, babies' good statistical intuitions are observed in the context of experience-based paradigms, whereas adults' blunders have commonly been observed in one-shot, description-based probability tasks (e.g., the Linda problem, the engineer-lawyer problem, the maternity ward problem; Kahneman, 2011), in which experiential learning is neither required nor permitted. Similarly, Hogarth and Soyer (2011) found that "even the statistically naïve

achieved accurate probabilistic inferences after experiencing sequentially simulated outcomes, and many preferred this presentation format" (p. 434).

All this does not mean that experience is a panacea for reasoning biases. Finding out to what extent experienced-based learning can foster decisions consistent with normative and adaptive concerns may, however, lead to a fuller understanding of human rationality. It may also help to integrate the conflicting findings on statistical reasoning that have emerged from the "man as an intuitive statistician" (Peterson & Beach, 1967) and the "heuristics-and-biases" (Kahneman, 2011) research programs. The latter's experimental protocol typically involves description-based scenarios; the former rested primarily on experienced-based settings.

Future Inquiries

A number of promising directions for research on the description-experience gap are opening up (e.g., Fantino & Navarro, 2012; see also Ludvig, Madan, Pisklak, & Spetch, 2014, on comparisons of choice tendencies in humans and other animals in experiential paradigms). Beyond the ones we have already discussed, let us briefly outline three more. First, decisions from experience require exploration, that is, the pursuit of the unknown. The scope of this pursuit inevitably molds the choice options experienced. This means that a person's risk preferences may not be revealed merely in her choice but even earlier, in her search policy (see Hills et al., 2013; Wulff et al., 2015b). Ideally, future theories of experiential choice should aim to jointly model search and choice and their dependencies. The models currently available—for instance, the ACT-R-inspired instance-based learning model (IBL, Gonzalez & Dutt, 2011) or reinforcement learning accounts, such as the value updating model or delta-rule learning (Busemeyer & Myung, 1992; Frey, Mata, & Hertwig, 2015)—predict choice but leave the dynamics of search largely unexplained.

Another little explored dimension of experiential choice is the extent to which search and choice and the description-experience gap depend on the structure of the choice options and the external circumstances of choice. Many real-world choices involve more than two options (e.g., Hills et al., 2013); many options have numerous possible consequences or represent a continuous distribution of consequences; and search and choice are often subject to time constraints (e.g., Glöckner et al., 2016; Lejarraga, Hertwig, & Gonzalez, 2012). Future research may examine the perspective of an active information searcher, who learns about her world not only within but also across problems (e.g., about the typical structure of the choices available). Such metalearning moves another often-ignored aspect to center stage: the individual's expectations prior to entering the experiment. Studies on decisions from experience usually provide little introductory information about the world of choices to be encountered in terms of their structure or outcome ranges. Evidence presented here and elsewhere (e.g., Fox & Hadar, 2006) suggests that people build up a model of the world they are navigating and employ it to, for instance, curtail search and make it more efficient (see also Dayan & Berridge, 2014; Dougherty, Thomas, & Lange, 2010; Ostwald et al., 2015).

Another important issue concerns the combination of both types of information and learning modes: People sometimes enjoy concurrent access to experience and description when making choices (see also Fantino & Navarro, 2012). Take risk warnings, for instance. These commonly involve written, graphic, or symbolic descriptions—for example, health warnings about the risk of contracting lung cancer from smoking or catching sexually transmitted diseases from unprotected sex. But they do not necessarily operate in an experiential vacuum. People have sometimes experienced numerous "safe" encounters with a hazardous event before being warned (e.g., repeated episodes of unprotected sex without getting a sexually transmitted disease). Sometimes people receive the warning immediately after disaster has struck; sometimes they are blank slates with no experience at all. How description-based warning and, more generally, verbal and written risk communication shapes future behavior is likely to codepend on people's past and recent experience. The intricate interplay of the description of what is often a rare risk and of experience that has not (yet) encountered the risk may offer one key to better understanding why risk warnings are often inefficient (Barron, Leider, & Stack, 2008) or result in counterintuitive effects, such as increased tolerance for risk (Newell, Rakow, Yechiam, & Sambur, 2015).

Furthermore, the smart combination of descriptions of risks with *simulated* experience in virtual realities may prove a valuable tool for conveying transparent and persuasive risk information in domains such as financial investments (e.g., Bradbury, Hens, & Zeisberger, 2014; Hogarth & Soyer, 2015a,2015b; Kaufmann, Weber, & Haisley, 2012; see also Lejarraga, Woike, & Hertwig, 2016; Weiss-Cohen, Konstantinidis, Speekenbrink, & Harvey, 2016) and climate change—a domain in which many people have, as yet, been spared direct personal experience of dreadful outcomes (Weber & Stern, 2011). These examples suggest that the description-experience gap is not only of theoretical importance but also has myriad practical implications.

Beyond Risky Choice: Description-Experience Gaps in Other Domains?

Research on the description-experience gap in choice between monetary lotteries is of key importance because findings from the description-only paradigm "have formed the bedrock of contemporary decision theories, most notably prospect theory" (Fantino & Navarro, 2012, p. 303). The impact of the description-experience distinction is, however, very likely not limited to choices between monetary lotteries. Numerous other choice and judgment phenomena have, for several decades, been studied primarily with description-based paradigms, including base-rate neglect, sunk-cost effects, and social and strategic dilemmas (see Fantino & Navarro, 2012). Recently, researchers have begun to examine the possibility of description-experience gaps in other domains, such as temporal discounting (Dai, Pachur, Pleskac, & Hertwig, 2017; Kemel & Travers, 2016), strategic reasoning in social games (Fleischhut, Artinger, Olschewski, Volz, & Hertwig, 2014; Martin, Gonzalez, Juvina, & Lebiere, 2014), consumer choice (Wulff, Hills, & Hertwig, 2015a), medical decisions and reasoning (Armstrong & Spaniol, 2017; Fraenkel, Peters, Tyra, & Oelberg, 2016; Lejarraga, Pachur, Frey, & Hertwig, 2015), and adolescent risk taking (Pollak et al., 2016; Rosenbaum, Venkatraman, Steinberg, & Chein, 2016; van den Bos & Hertwig, 2017).

Furthermore, description-experience gaps are likely to exist beyond judgment and choice phenomena. Consider for illustration causal reasoning. A normative causal theory such as causal Bayes

nets captures many aspects of human causal reasoning that sets it apart from purely associative and noncausal reasoning (see Glymour, 2003). Yet, some signature properties of causal Bayes nets are commonly violated—for example, the *explaining away* principle, according to which the presence of one cause in a common effect network makes another cause less likely. Within research on causal reasoning, two experimental methodologies have been commonly employed—one that experientially conveys the relevant statistical information and another that employs verbal descriptions of scenarios to convey causal models and the relevant parameter strengths. Rehder and Waldman (2017) systematically compared tasks in which causal scenarios were described (in terms of verbal statements of the causal relations) versus experienced (i.e., through samples of data representing the correlations implied by the causal relations). Their key finding was "stronger deviations from normative predictions [e.g., the explaining away principle] in the described conditions that highlight the instructed causal model compared to those that presented data [the experience condition]" (p. 1).

To conclude, description and experience are powerful modes of learning about environments far beyond the domain of risk choice. It is worth considering how these two modes of learning result in correlated or systematically different conclusions about human performance and even rationality. The description-experience gap represents a new point of entry for research on numerous cognitive functions. It holds the promise, when combined with a "do-it-both-ways" heuristic (Hertwig & Ortmann, 2001), of rapid progress in understanding the psychology and rationality of description and experience. Yet, let us preempt a potential misunderstanding: Description and experience should not be read as being in binary opposition. Descriptions come in many forms, as does experience, and sometimes they co-occur. Moreover, descriptions are not the only contrast to experiences and vice versa (see Jarvstad et al., 2013; Zeigenfuse et al., 2014). The general point is that cognitive functions and cognitive phenomena including reasoning, judgment, and choice may change, perhaps even substantially, as a function of the available mode of learning and representation format. This is a rich territory to explore.

References

*Studies included in the meta-analysis.

Abdellaoui, M., L'Haridon, O., & Paraschiv, C. (2011). Experienced vs. described uncertainty: Do we need two prospect theory specifications? *Management Science, 57,* 1879–1895. http://dx.doi.org/10.1287/mnsc.1110.1368

Allais, M. (1953). Le comportement de l'homme rationel devant le risque, critique des postulates et axiomes de l'école americaine. *Econometrica, 21,* 503–546. http://dx.doi.org/10.2307/1907921

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mathway, NJ: Erlbaum. http://dx.doi.org/10.4324/9781315805696

Anderson, N. H., & Hovland, C. I. (1957). The representation of order effects in communication research. In C. I. Hovland (Ed.), *The order of presentation in persuasion* (pp. 158–169). New Haven, CT: Yale University Press.

Armstrong, B., & Spaniol, J. (2017). Experienced probabilities increase understanding of diagnostic test results in younger and older adults. *Medical Decision Making*. Advance online publication. http://dx.doi.org/10.1177/0272989x17691954

Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica, 19,* 404–437. http://dx.doi.org/10.2307/1907465

Ashby, N. J., & Rakow, T. (2014). Forgetting the past: Individual differences in recency in subjective valuations from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1153–1162. http://dx.doi.org/10.1037/a0036352

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation, 2,* 89–195. http://dx.doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Clarendon Press.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16,* 215–233. http://dx.doi.org/10.1002/bdm.443

Barron, G., Leider, S., & Stack, J. (2008). The effect of safe experience on a warnings' impact: Sex, drugs, and rock-n-roll. *Organizational Behavior and Human Decision Processes, 106,* 125–142. http://dx.doi.org/10.1016/j.obhdp.2007.11.002

Barron, G., & Ursino, G. (2013). Underweighting rare events in experience-based decisions: Beyond sample error. *Journal of Economic Psychology, 39,* 278–286. http://dx.doi.org/10.1016/j.joep.2013.09.002

Barron, G., & Yechiam, E. (2009). The coexistence of overestimation and underweighting of rare events and the contingent recency effect. *Judgment and Decision Making, 4,* 447–460.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica: Journal of the Econometric Society, 22,* 23–36. (Original work published 1738)

Borgenstein, M. H., Hedges, L. V., Higgins, L., & Rothstein, J. P. T. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470743386

Bradbury, M. A., Hens, T., & Zeisberger, S. (2014). Improving investment decisions with simulated experience. *Review of Finance, 19,* 1019–1052. http://dx.doi.org/10.1093/rof/rfu021

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113,* 409–432. http://dx.doi.org/10.1037/0033-295X.113.2.409

Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision, 1,* 252–274. http://dx.doi.org/10.1037/dec0000020

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539–576. http://dx.doi.org/10.1037/0033-295X.114.3.539

Budescu, D. V., & Wallsten, T. S. (1987). Subjective estimation of precise and vague uncertainties. In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 63–81). Chichester, UK: Wiley.

Busemeyer, J. R. (1982). Choice behavior in a sequential decision-making task. *Organizational Behavior and Human Performance, 29,* 175–207. http://dx.doi.org/10.1016/0030-5073(82)90255-0

Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 538–564. http://dx.doi.org/10.1037//0278-7393.11.3.538

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121,* 177–194. http://dx.doi.org/10.1037/0096-3445.121.2.177

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. Oxford, UK: Wiley. http://dx.doi.org/10.1037/14496-000

*Camilleri, A. R., & Newell, B. R. (2009a). The role of representation in experience-based choice. *Judgment and Decision Making, 4,* 518–529.

*Camilleri, A. R., & Newell, B. R. (2009b). Within-subject preference reversals in description- and experience-based choice. In N. Taatgen & H. V. Rijn (Eds.), In *Proceedings of the 31st annual conference of the*

cognitive science society (pp. 449–454). Austin, TX: Cognitive Science Society.

*Camilleri, A. R., & Newell, B. R. (2011a). Description- and experience-based choice: Does equivalent information equal equivalent choice? *Acta Psychologica, 136,* 276–284. http://dx.doi.org/10.1016/j.actpsy.2010.11.007

*Camilleri, A. R., & Newell, B. R. (2011b). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review, 18,* 377–384. http://dx.doi.org/10.3758/s13423-010-0040-2

Camilleri, A. R., & Newell, B. R. (2013). The long and short of it: Closing the description-experience "gap" by taking the long-run view. *Cognition, 126,* 54–71. http://dx.doi.org/10.1016/j.cognition.2012.09.001

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 362,* 933–942. http://dx.doi.org/10.1098/rstb.2007.2098

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences, 24,* 154–176. http://dx.doi.org/10.1017/S0140525X0161392X

Dai, J., Pachur, T., Pleskac, T., & Hertwig, R. (2017). *A description-experience gap in intertemporal choice*. Manuscript submitted for publication.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience, 14,* 473–492. http://dx.doi.org/10.3758/s13415-014-0277-8

Denrell, J. C. (2015). Reference-dependent risk sensitivity as rational inference. *Psychological Review, 122,* 461–484. http://dx.doi.org/10.1037/a0039250

Dougherty, M., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychology of Learning and Motivation, 52,* 299–342. http://dx.doi.org/10.1016/S0079-7421(10)52008-5

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51,* 380–417. http://dx.doi.org/10.1037/h0053870

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4,* 59–74. http://dx.doi.org/10.1177/001872086200400201

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics, 75,* 643–669. http://dx.doi.org/10.2307/1884324

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*. http://dx.doi.org/10.1037/rev0000062

*Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23,* 15–47. http://dx.doi.org/10.1002/bdm.683

Erev, I., Glozman, I., & Hertwig, R. (2008). What impacts the impact of rare events. *Journal of Risk and Uncertainty, 36,* 153–177. http://dx.doi.org/10.1007/s11166-008-9035-z

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review, 88,* 848–881.

Estes, W. K. (1959). Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 9–52). Stanford, CA: Stanford University Press.

Fantino, E., & Navarro, A. (2012). Description-experience gaps: Assessments in other choice paradigms. *Journal of Behavioral Decision Making, 25,* 303–314. http://dx.doi.org/10.1002/bdm.737

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9,* 59–79. http://dx.doi.org/10.3758/BF03196257

*Fleischhut, N., Artinger, F., Olschewski, S., Volz, K. G., & Hertwig, R. (2014). Sampling of social information: Decisions from experience in bargaining. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Program of the 36th annual conference of the cognitive science society* (pp. 1048–1053). Austin, TX: Cognitive Science Society. http://dx.doi.org/10.1037/e528942014-510

Fox, C. R., & Hadar, L. (2006). "Decisions from experience" = sampling error plus prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making Journal, 1,* 159–161.

Fraenkel, L., Peters, E., Tyra, S., & Oelberg, D. (2016). Shared medical decision making in lung cancer screening: Experienced versus descriptive risk formats. *Medical Decision Making, 36,* 518–525. http://dx.doi.org/10.1177/0272989X15611083

*Frey, R., Hertwig, R., & Rieskamp, J. (2014). Fear shapes information acquisition in decisions from experience. *Cognition, 132,* 90–99. http://dx.doi.org/10.1016/j.cognition.2014.03.009

*Frey, R., Mata, R., & Hertwig, R. (2015). The role of cognitive abilities in decisions from experience: Age differences emerge as a function of choice set size. *Cognition, 142,* 60–80. http://dx.doi.org/10.1016/j.cognition.2015.05.004

Fried, L. S., & Peterson, C. R. (1969). Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology, 80,* 525–529. http://dx.doi.org/10.1037/h0027484

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of political Economy, 56,* 279–304. http://dx.doi.org/10.1086/256692

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis, 25,* 623–629. http://dx.doi.org/10.1111/j.1539-6924.2005.00608.x

*Glöckner, A., Fiedler, S., Hochman, G., Ayal, S., & Hilbig, B. E. (2012). Processing differences between descriptions and experience: A comparative analysis using eye-tracking and physiological measures. *Frontiers in Psychology, 3,* 173. http://dx.doi.org/10.3389/fpsyg.2012.00173

*Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General, 145,* 486–508. http://dx.doi.org/10.1037/a0040103

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition, 123,* 21–32. http://dx.doi.org/10.1016/j.cognition.2011.12.002

Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences, 7,* 43–48. http://dx.doi.org/10.1016/S1364-6613(02)00009-8

Goldstein, W. M., & Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. *Psychological Review, 94,* 236–254. http://dx.doi.org/10.1037//0033-295x.94.2.236

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118,* 523–551. http://dx.doi.org/10.1037/a0024558

*Gonzalez, C., & Mehlhorn, K. (2015). Framing from experience: Cognitive processes and predictions of risky choice. *Cognitive Science*.

Gopnik, A. (2014, January 10). The surprising probability gurus wearing diapers. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/SB10001424052702303393804579308662389246416

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7,* 464–481. http://dx.doi.org/10.1177/1745691612454304

Hacking, I. (1975). *The emergence of probability*. Cambridge, UK: New York: Cambridge University Press. http://dx.doi.org/10.1017/s0031819100018866

*Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making, 4,* 317–325.

Harman, J. L., & Gonzalez, C. (2015). Allais from experience: Choice consistency, rare events, and common consequences in repeated decisions. *Journal of Behavioral Decision Making, 28,* 369–381. http://dx.doi.org/10.1002/bdm.1855

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388. http://dx.doi.org/10.1037/0003-066X.39.12.1372

Hau, R., Pleskac, T. J., & Hertwig, R. (2010). Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making, 23,* 48–68. http://dx.doi.org/10.1002/bdm.665

*Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518. http://dx.doi.org/10.1002/bdm.598

Hawkins, G., Camilleri, A., Heathcote, A., Newell, B., & Brown, S. (2014). Modeling probability knowledge and choice in decisions from experience. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese, 187,* 269–292. http://dx.doi.org/10.1007/s11229-011-0024-4

Hertwig, R. (2016). Decisions from experience. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 239–267). Oxford, UK: Blackwell Wiley. http://dx.doi.org/10.1002/9781118468333.ch8

*Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511614576.004

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13,* 517–523. http://dx.doi.org/10.1016/j.tics.2009.09.004

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24,* 383–403. http://dx.doi.org/10.1037/e683322011-032

Hertwig, R., & Pleskac, T. J. (2008). The game of life: How small samples render choice simpler. In N. Chater & M. Oaksford (Ed.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 209–235). Retrieved from http://dx.doi.org/10.1093/acprof:oso/9780199216093.003.0010

*Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115,* 225–237. http://dx.doi.org/10.1016/j.cognition.2009.12.009

Hilbig, B. E., & Glöckner, A. (2011). Yes, they can! Appropriate weighting of small probabilities as a function of information acquisition. *Acta Psychologica, 138,* 390–396. http://dx.doi.org/10.1016/j.actpsy.2011.09.005

Hilgard, E. R., & Marquis, D. G. (1940). *Conditioning and learning*. New York, NY: Appleton. http://dx.doi.org/10.1037/14591-002

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience. Do our patterns of sampling foreshadow our decisions? *Psychological Science, 21,* 1787–1792. http://dx.doi.org/10.1177/0956797610387443

*Hills, T. T., Noguchi, T., & Gibbert, M. (2013). Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychonomic Bulletin & Review, 20,* 1023–1031. http://dx.doi.org/10.3758/s13423-013-0422-3

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: kind experience versus nontransparent description. *Journal of Experimental Psychology: General, 140,* 434. http://dx.doi.org/10.1037/a0023265

Hogarth, R. M., & Soyer, E. (2015a). Communicating forecasts: The simplicity of simulated experience. *Journal of Business Research, 68,* 1800–1809. http://dx.doi.org/10.1016/j.jbusres.2015.03.039

Hogarth, R. M., & Soyer, E. (2015b). Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory & Cognition, 4,* 221–228. http://dx.doi.org/10.1016/j.jarmac.2014.01.005

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review, 92,* 1644–1655. http://dx.doi.org/10.1257/000282802762024700

Jarvstad, A., Hahn, U., Rushton, S. K., & Warren, P. A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 16271–16276. http://dx.doi.org/10.1073/pnas.1300239110

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19,* 1015–1022. http://dx.doi.org/10.1111/j.1467-9280.2008.02193.x

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Macmillan. http://dx.doi.org/10.5840/inquiryct201227212

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291. http://dx.doi.org/10.2307/1914185

Kaufmann, C., Weber, M., & Haisley, E. C. (2012). The role of experience sampling and graphical displays on one's investment risk appetite. *Management Science, 59,* 323–340. http://dx.doi.org/10.1287/mnsc.1120.1607

*Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in)variant are subjective representations of described and experienced risk and rewards? *Cognition, 157,* 126–138. http://dx.doi.org/10.1016/j.cognition.2016.08.020

Kemel, E., & Travers, M. (2016). Comparing attitudes toward time and toward money in experience-based decisions. *Theory and Decision, 80,* 71–100. http://dx.doi.org/10.1007/s11238-015-9490-3

Knight, F. H. (1921). *Risk, uncertainty and profit*. New York, NY: Hart, Schaffner and Marx. http://dx.doi.org/10.1017/cbo9780511817410.005

Kudryavtsev, A., & Pavlodsky, J. (2012). Description-based and experience-based decisions: Individual analysis. *Judgment and Decision Making, 7,* 316–331.

Lee, W. (1971). *Decision theory and human behavior*. New York, NY: Wiley.

*Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making, 23,* 100–116. http://dx.doi.org/10.1002/bdm.666

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25,* 143–153. http://dx.doi.org/10.1002/bdm.722

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes, 116,* 286–295. http://dx.doi.org/10.1016/j.obhdp.2011.05.001

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition, 124,* 334–342. http://dx.doi.org/10.1016/j.cognition.2012.06.002

Lejarraga, T., & Müller-Trede, J. (2016). When experience meets description: How dyads integrate experiential and descriptive information in risky decisions. *Management Science*. Advance online publication.

*Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2015). Decisions from experience: From monetary to medical gambles. *Journal of Behavioral Decision Making*. Advance online publication.

Lejarraga, T., Woike, J. K., & Hertwig, R. (2016). Description and experience: How experimental investors learn about booms and busts affects their financial risk taking. *Cognition, 157,* 365–383. http://dx.doi.org/10.1016/j.cognition.2016.10.001

Light, R. J., Singer, J., & Willett, J. (1994). The Visual Presentation and Interpretation of Meta-Analyses. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.

Lin, D., Donkin, C., & Newell, B. (2015). The exemplar confusion model: An account of biased probability estimates in decisions from description. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 1374–1379). Austin, TX: Cognitive Science Society.

Lindskog, M., Winman, A., & Juslin, P. (2013). Calculate or wait: Is man an eager or a lazy intuitive statistician? *Journal of Cognitive Psychology, 25,* 994–1014. http://dx.doi.org/10.1080/20445911.2013.841170

Luce, R. D., & Raiffa, H. (2012). *Games and decisions: Introduction and critical survey*. New York, NY: Courier Corporation. (Original work published 1957)

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 3, pp. 249–410). New York, NY: Wiley.

Ludvig, E. A., Madan, C. R., Pisklak, J. M., & Spetch, M. L. (2014). Reward context determines risky choice in pigeons and humans. *Biology Letters*. Advance online publication. http://dx.doi.org/10.1098/rsbl.2014.0451

Ludvig, E. A., & Spetch, M. L. (2011). Of black swans and tossed coins: Is the description-experience gap in risky choice limited to rare events? *PLoS ONE, 6,* e20262. http://dx.doi.org/10.1371/journal.pone.0020262

March, J. G. (1996). Learning to be risk averse. *Psychological Review, 103,* 309–319. http://dx.doi.org/10.1037/0033-295X.103.2.309

March, J. G., & Olsen, J. P. (2010). *Rediscovering institutions*. New York, NY: Simon & Schuster.

Markant, D. B., Pleskac, T. J., Diederich, A., Pachur, T., & Hertwig, R. (2015). Modeling choice and search in decisions from experience: A sequential sampling approach. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 1512–1517). Austin, TX: Cognitive Science Society.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy, 60,* 151–158. http://dx.doi.org/10.1086/257177

Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (2014). A description-experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making, 27,* 349–362. http://dx.doi.org/10.1002/bdm.1810

*Mehlhorn, K., Ben-Asher, N., Dutt, V., & Gonzalez, C. (2014). Observed variability and values matter: Toward a better understanding of information search and decisions from experience. *Journal of Behavioral Decision Making, 27,* 328–339. http://dx.doi.org/10.1002/bdm.1809

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

Mullett, T. L., & Stewart, N. (2016). Implications of visual attention phenomena for models of preferential choice. *Decision, 3,* 231–253. http://dx.doi.org/10.1037/dec0000049

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488. http://dx.doi.org/10.1037/h0045106

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116,* 499–518. http://dx.doi.org/10.1037/a0016104

Newell, B. R., Rakow, T., Yechiam, E., & Sambur, M. (2015). Rare disaster information can increase risk-taking. *Nature Climate Change, 6,* 158–161. http://dx.doi.org/10.1038/nclimate2822

Noguchi, T., & Hills, T. T. (2016). Description-experience gap in choice deferral. *Decision, 3,* 54–61. http://dx.doi.org/10.1037/dec0000044

*Noguchi, T., & Hills, T. T. (2016). Experience-based decisions favor riskier alternatives in large sets. *Journal of Behavioral Decision Making, 29,* 489–498. http://dx.doi.org/10.1002/bdm.1893

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin, 142,* 758–799. http://dx.doi.org/10.1037/bul0000046

Ostwald, D., Starke, L., & Hertwig, R. (2015). A normative inference approach for optimal sample sizes in decisions from experience. *Frontiers in Psychology, 6,* 1342. http://dx.doi.org/10.3389/fpsyg.2015.01342

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46. http://dx.doi.org/10.1037/h0024722

*Phillips, N. D., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition, 133,* 104–119. http://dx.doi.org/10.1016/j.cognition.2014.06.006

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. London, UK: Penguin. http://dx.doi.org/10.5334/opt.070912

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General, 143,* 2000–2019. http://dx.doi.org/10.1037/xge0000013

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review, 122,* 621–647. http://dx.doi.org/10.1037/a0039413

Pollak, Y., Oz, A., Neventsal, O., Rabi, O., Kitrossky, L., & Maeir, A. (2016). Do adolescents with attention-deficit/hyperactivity disorder show risk seeking? Disentangling probabilistic decision making by equalizing the favorability of alternatives. *Journal of Abnormal Psychology, 125,* 387–398. http://dx.doi.org/10.1037/abn0000140

Prelec, D. (1998). The probability weighting function. *Econometrica, 66,* 497–527. http://dx.doi.org/10.2307/2998573

*Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106,* 168–179. http://dx.doi.org/10.1016/j.obhdp.2008.02.001

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making, 23,* 1–14. http://dx.doi.org/10.1002/bdm.681

*Rakow, T., & Rahim, S. B. (2010). Developmental insights into experience-based decision making. *Journal of Behavioral Decision Making, 23,* 69–82. http://dx.doi.org/10.1002/bdm.672

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rehder, B., & Waldman, M. R. (2017). *Failures of explaining away and screening off in described versus experienced causal learning scenarios*. *Memory & Cognition, 45,* 245–260.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts. http://dx.doi.org/10.1007/978-3-540-29678-2_5982

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review, 77,* 481–495. http://dx.doi.org/10.1037/h0029964

Rosenbaum, G. M., Venkatraman, V., Steinberg, L., & Chein, J. M. (2016). *The influences of described and experienced information on adolescent risky decision-making*. Manuscript submitted for publication.

Savage, L. J. (1954/1972). *The foundations of statistics*. New York, NY: Dover Publications.

Schmandt-Besserat, D. (1996). *How writing came about*. Austin, TX: University of Texas Press. http://dx.doi.org/10.1075/wll.1.2.08eng

Schulze, C., & Hertwig, R. (2017). *Statistical intuitions: Smart babies, stupid adults?* Manuscript submitted for publication.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience, 6,* 1317–1322. http://dx.doi.org/10.1038/nn1150

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General, 139,* 266–298. http://dx.doi.org/10.1037/a0018620

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MY: MIT Press.

Taleb, N. N. (2010). *The black swan: The impact of the highly improbable fragility* (2nd ed.). New York, NY: Random House.

Trautmann, S. T., & van de Kuilen, G. (2016). Ambiguity attitudes. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (Vol. 1, pp. 89–116). Oxford, England: Blackwell Wiley. http://dx.doi.org/10.1002/9781118468333.ch3

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review, 102,* 269–283. http://dx.doi.org/10.1037/0033-295X.102.2.269

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323. http://dx.doi.org/10.1007/BF00122574

*Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science, 20,* 473–479. http://dx.doi.org/10.1111/j.1467-9280.2009.02319.x

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505. http://dx.doi.org/10.3758/BF03192720

van de Kuilen, G., & Wakker, P. P. (2011). The midweight method to measure attitudes toward risk and ambiguity. *Management Science, 57,* 582–598. http://dx.doi.org/10.1287/mnsc.1100.1282

van den Bos, W., & Hertwig, R. (2017). Adolescents display distinctive tolerance to ambiguity and to uncertainty during risky decision making. *Scientific Reports*. http://dx.doi.org/10.1038/srep40962

van den Bos, W., Jenny, M. A., & Wulff, D. U. (2014). Open minded psychology. In S. A. Moore (Ed.), *Issues in open research data* (pp. 107–127). London, UK: Ubiquity Press. http://dx.doi.org/10.5334/ban.g

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. http://dx.doi.org/10.18637/jss.v036.i03

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38,* 599–637. http://dx.doi.org/10.1111/cogs.12101

Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111,* 430–445. http://dx.doi.org/10.1037/0033-295X.111.2.430

Weber, E. U., & Stern, P. C. (2011). Public understanding of climate change in the United States. *American Psychologist, 66,* 315–328. http://dx.doi.org/10.1037/a0023253

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale: Revised manual (WAIS-R)*. New York, NY: Psychological Corporation.

Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes, 135,* 55–69. http://dx.doi.org/10.1016/j.obhdp.2016.05.005

*Wulff, D., & Hertwig, R. (2012). *Replication of Hertwig, Barron, Weber, & Erev (2004)*. Unpublished raw data.

*Wulff, D. U., Hills, T. T., & Hertwig, R. (2015a). Online product reviews and the description-experience gap. *Journal of Behavioral Decision Making, 28,* 214–223. http://dx.doi.org/10.1002/bdm.1841

*Wulff, D. U., Hills, T. T., & Hertwig, R. (2015b). How short- and long-run aspirations impact search and choice in decisions from experience. *Cognition, 144,* 29–37. http://dx.doi.org/10.1016/j.cognition.2015.07.006

Wulff, D. U., & Pachur, T. (2016). Modeling valuations from experience: A comment on Ashby and Rakow (2014). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 158–166. http://dx.doi.org/10.1037/xlm0000165

Wulff, D. U., & van den Bos, W. (2017). Modeling delay discounting. *Psychological Science*.

Yechiam, E., Barron, G., & Erev, I. (2005). The role of personal experience in contributing to different patterns of response to rare terrorist attacks. *The Journal of Conflict Resolution, 49,* 430–439. http://dx.doi.org/10.1177/0022002704270847

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *ETC frequency processing and cognition* (pp. 21–36). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198508632.003.0002

Zeigenfuse, M. D., Pleskac, T. J., & Liu, T. (2014). Rapid decisions from experience. *Cognition, 131,* 181–194. http://dx.doi.org/10.1016/j.cognition.2013.12.012

(*Appendices follow*)

Appendix A

Table A1
Sampling Paradigm Data Sets Used in the Analysis (Ordered by Year of Publication)

Article	Short	No.	*N*	Problems	*N* outcomes	Certain	Problems/ Ppt	Type	Sample size	*N* description	Within	Incentive	Note
Hertwig et al. (2004)	HBWE04	1	50	4G, 2L	4	4	3	Autonom.	15	50	No	Yes	—
Hau et al. (2008)	HPKH08	1	42	4G, 2L	4	4	5	Autonom.	11	109	No	Yes	S1, HBWE04
		2	39	4G, 2L	4	4	6	Autonom.	33	—	No	Yes	replication S2,
		3	40	4G, 2L	4	4	6	Regulated	100	—	No	No	High incentives S3
Rakow et al. (2008)	RDN08	1	80	9G, 3L	4	9	6	Autonom.	15	80	No	Yes	—
Camilleri and Newell (2009a)	CN09a	1	80	4G, 4L	3	8	2	Matched	25	80	Yes	Yes	Frequency judgment after choice
		2	80	4G, 4L	3	8	2	Matched	26	80	Yes	Yes	Frequency judgment before choice
Camilleri and Newell (2009b)	CN09b	1	20	7G, 3L	3	10	10	Autonom.	7.5	20	Yes	Yes	—
	7G, 3L	3	10	10				Autonom.	10	20	Yes	Yes	Description first
Hadar and Fox (2009)	HF09	1	23	2G, 1L	4	2	3	Regulated	20	—	No	No	Outcomes disclosed
		2	31	2G, 1L	4	2	3	Regulated	20	—	No	No	Shapes as outcomes
		3	30	2G, 1L	4	2	3	Regulated	20	—	No	No	—
		4	27	2G, 1L	4	2	3	Regulated	20	—	No	No	Outcomes disclosed, shapes as outcomes
Ungemach et al. (2009)	UCS09	1	25	4G, 2L	4	4	6	Autonom.	19	25	No	No	S1
		2	25	4G, 2L	4	4	6	Regulated	19	—	No	No	S1, Yoked design
		3	197	4G, 2L	4	4	1	Regulated	80	—	No	No	S2
Erev et al. (2010)	TPT10	1	39	20G, 20L, 20M	3	60	30	Autonom.	9	20	No	Yes	Estimation set
		2	40	20G, 20L, 20M	3	60	30	Autonom.	11	20	No	Yes	Competition set
Hertwig and Pleskac (2010)	HP10	1	88	8G, 4L	4	8	12	Autonom.	10	89	No	Yes	—
Lejarraga (2010)	L10	1	85	4G, 3L	4	3	3	Autonom.	37	83	Yes	Yes	Self-selected format
Rakow and Rahim (2010)	RR10	1	26	4G	3	4	4	Regulated	20	26	No	No	S1, 5–6 years old
		2	25	4G	3	4	4	Regulated	20	24	No	No	S1, adults
		3	38	6G	3	6	6	Regulated	20	37	No	No	S2, 5–6 years old
		4	37	6G	3	6	6	Regulated	20	40	No	No	S2, 16–17 years old
		5	40	6G	3	6	6	Regulated	20	40	No	No	S2, 16–17 years old, descr. first
		6	19	2G, 4M	3	6	6	Regulated	20	19	Yes	No	S3, 12–13 years old
		7	20	2G, 4M	3	6	6	Regulated	20	20	Yes	No	S3, 16–17 years old
		8	17	2G, 4M	3	6	6	Regulated	20	17	Yes	No	S3, 12–13 years old, descr. first
		9	17	2G, 4M	3	6	6	Regulated	20	17	Yes	No	S3, 16–17 years old, descr. first
Camilleri (2011a)	CN11a	1	40	2G, 2L	3	4	4	Regulated	100	40	No	Yes	—
Camilleri (2011b)	CN11b	1	31	7G, 3L	3	10	10	Matched	10	36	No	Yes	Pseudo-random
		2	35	7G, 3L	3	10	10	Matched	10	—	No	Yes	Pseudo-random
		3	36	5G, 3L	3	8	8	Regulated	20	—	No	Yes	—
Glöckner et al. (2012)	GFHAH12	1	22	37G	4	0	37	Autonom.	30	22	No	Yes	Eye-tracker, target problems
		2	22	22G	4	0	22	Autonom.	27	22	No	Yes	Eye-tracker, filler problems
Wulff and Hertwig (2012)	WH12	1	59	4G, 2L	4	4	6	Autonom.	13	59	Yes	Yes	MTurk, World
		2	78	4G, 2L	4	4	6	Autonom.	10	78	Yes	Yes	MTurk, World, descr. first
		3	41	4G, 2L	4	4	6	Autonom.	13	41	Yes	Yes	MTurk, US only
		4	40	4G, 2L	4	4	6	Autonom.	9	40	Yes	Yes	MTurk, US only, descr. first
Hills et al. (2013)	HNG13	1	32	1G	4	0	1	Autonom.	4.5	—	No	Yes	Two-to-many options
		2	32	1G	4	0	1	Autonom.	5.5	—	No	Yes	Many-to-two options
Fleischhut et al. (2014)	FAOVH14	1	46	12G	4	4	12	Autonom.	22	47	No	Yes	—
		2	43	12G	4	4	12	Regulated	40	—	No	Yes	—

(*Appendices continue*)

Table A1 (*cotinued*)

Article	Short	No.	*N*	Problems	*N* outcomes	Certain	Problems/ Ppt	Type	Sample size	*N* description	Within	Incentive	Note
Frey et al. (2014)	FHR14	1	27	5G, 4L	4	5	9	Autonom.	28	—	No	Yes	S1, Mood: Happy
		2	28	5G, 4L	4	5	9	Autonom.	32	—	No	Yes	S1, Mood: Sad
		3	29	5G, 4L	4	5	9	Autonom.	44	—	No	Yes	S1, Mood: Fearful
		4	28	5G, 4L	4	5	9	Autonom.	32	—	No	Yes	S1, Mood: Angry
		5	23	2G, 2L	3	4	4	Autonom.	40	—	No	Yes	S2, Dental surgeon
		6	26	2G, 2L	3	4	4	Autonom.	4	—	No	Yes	S2, Comedy show
Mehlhorn et al. (2014)	MBDG14	1	294	8G, 8L	3	16	2	Autonom.	4	—	No	Yes	MTurk
Phillips et al. (2014)	PHKA14	1	36	21M	4	0	5	Autonom.	19	—	No	Yes	—
Frey et al. (2015)	FMH15	1	60	6G, 6L	4	3	12	Autonom.	24	—	No	Yes	S1, Younger adults
		2	61	6G, 6L	4	3	12	Autonom.	17	—	No	Yes	S1, Older adults
		3	35	100G, 100L	6	0	84	Autonom.	21	—	No	Yes	S2, Younger adults
		4	35	100G, 100L	6	0	85	Autonom.	18	—	No	Yes	S2, Older adults
Gonzalez and Mehlhorn (2015)	GM15	1	125	1G, 1L	3	2	1	Autonom.	4	102	No	No	S1, MTurk, Asian disease problem
		2	400	2G, 2L	3	4	1	Regulated	5	—	No	No	(ADP) S2, MTurk, ADP, 5 samples S2, MTurk, ADP, 100 samples
		3	400	2G, 2L	3	4	1	Regulated	100	—	No	No	
Noguchi and Hills (2015)	NH15	1	32	185G, 187L	4	0	6	Autonom.	5	32	No	Yes	MTurk
		2	21	117G, 120L	4	0	6	Autonom.	3	20	No	Yes	Lab
Wulff et al. (2015a)	WHH15a	1	30	8G	10	0	8	Autonom.	20	30	Yes	Yes	Consumer ratings
		2	33	8G	10	0	8	Autonom.	16	33	Yes	Yes	Consumer ratings, descr. first
Wulff et al. (2015b)	WHH15b	1	41	16G	4	9	16	Autonom.	19	—	No	Yes	Single-play
		2	42	16G	4	9	16	Autonom.	22	—	No	Yes	Multiplay
		3	41	16G	4	9	16	Autonom.	20	—	No	Yes	—
Glöckner et al. (2016)	GHHF16	1	28	40G	4	0	39	Autonom.	30	24	No	Yes	S1, target problems
		2	28	11G	4	0	11	Autonom.	27	24	No	Yes	S1, filler problems
		3	28	9G	4	7	8	Autonom.	39	24	No	Yes	S1, H/CN problems
		4	24	38G	4	0	38	Autonom.	41	53	No	Yes	S2, Target problems
		5	24	22G	4	0	22	Autonom.	38	53	No	Yes	S2, Filler problems
		6	25	38G	4	0	37	Autonom.	34	—	No	Yes	S3, Outcomes disclosed, target problems
		7	25	22G	4	0	22	Autonom.	27.5	—	No	Yes	S3, Outcomes disclosed, filler problems
		8	18	35G, 15L, 19M	4	3	69	Autonom.	42	18	No	Yes	S4, Problem set A
		9	20	35G, 15L, 19M	4	0	67	Autonom.	44	18	No	Yes	S4, Problem set B
Kellen et al. 2016)	KPH16	1	104	6G, 6L	4	3	12	Autonom.	18	104	Yes	Yes	HWBE04/HP10
		2	104	22G, 22L, 22M	4	0	66	Autonom.	19	104	Yes	Yes	problems Random problems
		3	104	10G, 10L, 8M	4	10	28	Autonom.	17	104	Yes	Yes	Loss/risk problems
		4	104	4G, 4M	4	6	8	Autonom.	18	104	Yes	Yes	Other problems
Lejarraga et al. (2016)	LPFH16	1	30	4G, 440L	4	53	9	Autonom.	19	30	No	No	Monetary problems
		2	30	4G, 440L	4	53	9	Autonom.	15	30	No	No	Medical problems

Note. Short = Study abbreviation used in the Figures. *N* = Number of participants in the unit. Problems = Number of problems using gain (G), loss (L), and mixed (M) decision problems. Certain = Number of problems containing a sure-event option. *N* outcomes = The maximum number of outcomes across both options. Autonom. = Autonomous sampling. Descr. first = Description first task in study, otherwise experience first. S1, S2, S3 = Study 1, Study 2, Study 3.

(*Appendices continue*)

Appendix B

Publication Bias

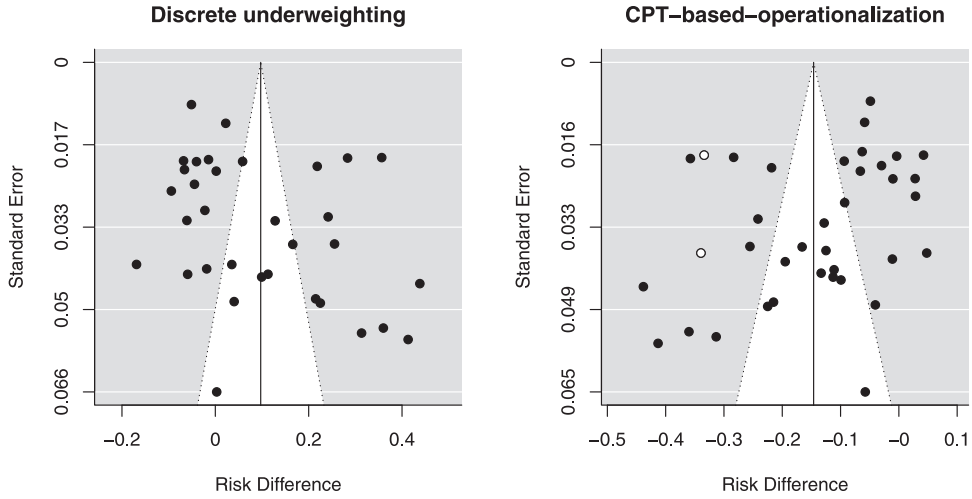

Figure B1. Funnel plots of the description-experience gap in the sampling paradigm (see Figure 2). Plots show the results after imputing missing values (white dots) using the trim and fill method. The vertical line reflects the pooled mean effect size after trim and fill correction. The diagonal lines represent the corresponding 95% confidence intervals.

Appendix C

The Description-Experience Gap in the Partial-Feedback Paradigm

In order to provide a point of reference for the present analysis of the sampling paradigm, we also conducted a (limited) literature search on the description-experience gap in the partial-feedback paradigm. Specifically, we screened all 229 citations of Barron and Erev (2003), the classic article that established the description-experience gap in the partial-feedback design. Of these, 152 presented original data, 52 required participants to make repeated, nonsocial choices between payoff distributions, and seven contained comparable decisions from description data. As we relied on standard meta-analysis techniques for this data, we refrained from requesting the original data. See Table C1 for details of the studies included.

Comparison of the two paradigms suggested that the average description-experience gap was smaller in the sampling paradigm than in the partial-feedback paradigm: 9.7 and −13.4 percentage points versus 21.4 and −18.6 percentage points, as measured by the discrete and CPT-based operationalizations, respectively (see Table C2 and C3; see also Figure C1). One explanation for this difference may be problem structure. For instance, 86% of the 14,421 choices considered in the partial-feedback paradigm were between a risky and a safe option, whereas this applied to only 28% of choices in the sampling paradigm. As Table C3 shows, the gap in the partial-feedback paradigm appears to be subject to the same moderators as the gap in the sampling paradigm, including the problem structure and the probability of the rarest event. When these are accounted for by, for instance, considering only certain versus risky choices with rare events of $p < .15$ (i.e., the decision problems for which the gap is largest in both paradigms), both paradigms produce comparable gap sizes of more than 20 percentage points.

(*Appendices continue*)

Table C1
Partial-Feedback Paradigm Studies Used in the Analysis (Ordered by Year of Publication)

Article	*N* Experience	*N* Description	Problems	Certain	Notes
Barron and Erev (2003)	24	91	4G, 1L	3	
Yechiam et al. (2005)	24	30	1L	0	
Erev et al. (2010)	20	20	20G, 20L, 20M	60	Estimation set
	20	20	20G, 20L, 20M	60	Prediction set
Camilleri and Newell (2011b)	40	40	2G, 2L	4	
Lejarraga and Gonzalez (2011)	31	30	4G	4	
Kudryavtsev and Pavlodsky (2012)	75	75	10M	0	Three-outcome problems
Camilleri and Newell (2013)	102	102	16G, 16L		
Harman and Gonzalez (2015)	100	100	2G	0	Allais's paradox problem (Allais, 1953)

Table C2
The Description-Experience Gap in the Partial-Feedback Paradigm

Study	Discrete underweighting			CPT		
	Experience	Description	Gap	Experience	Description	Gap
Barron and Erev (2003)	.54	.41	.13	.37	.58	−.21
Yechiam et al. (2005)	.69	.40	.29	.31	.60	−.29
Erev et al. (2010), Exp. 1	.59	.30	.29	.47	.70	−.22
Erev et al. (2010), Exp. 2	.59	.27	.31	.46	.77	−.31
Camilleri and Newell (2011a)	.76	.37	.39	.24	.63	−.39
Lejarraga and Gonzalez (2011)	.76	.54	.22	.24	.47	−.22
Kudryavtsev and Pavlodsky (2012)	.45	.41	.05	.29	.21	.08
Camilleri and Newell (2013)	.52	.54	−.02	.48	.46	.02
Harman and Gonzalez (2015)	.49	.26	.22	.55	.66	−.10
Weighted average[a] (standard error)			.214 (.02)			−.186 (.02)

[a] Based on a meta-analytical random effects model (see Borgenstein, Hedges, Higgins, and Rothstein (2009)).

Table C3
The Description-Experience Gap as a Function of the Structure of the Choice and the Probability of the Rarest Event (in Percentage Points)

Structure	Discrete underweighting			CPT		
	Rare event < .15	Rare event ≥ .15	Marginal	Rare event < .15	Rare event ≥ .15	Marginal
Sampling paradigm						
Certain vs. risky	**23.5 (3.5)**	**17.1 (3.6)**	**19.7 (3.0)**	**−23.5 (3.4)**	**−18.0 (2.8)**	**−20.1 (2.5)**
Risky vs. risky	**6.7 (2.8)**	−2.7 (1.9)	2.1 (1.7)	**−12.4 (2.3)**	−2 (1.5)	**−7.5 (1.4)**
Marginal	**13.8 (3.1)**	**7.6 (2.0)**	**9.7 (2.6)**	**−16.9 (2.4)**	**−10.0 (1.6)**	**−13.4 (1.9)**
Partial-feedback paradigm						
Certain vs. risky	**27.6 (2.2)**	−4.3 (3.0)	**22.8 (2.3)**	**−24.6 (2.3)**	2.0 (5.5)	**−20.2 (2.3)**
Risky vs. risky	**13.0 (4.1)**	−2 (5.5)	**7.9 (3.7)**	−3.8 (6.0)	4.3 (2.9)	−1.1 (4.2)
Marginal	**26.4 (2.0)**	−2.3 (4.7)	**21.4 (2.1)**	−23.1 (2.2)	2.3 (4.8)	**−18.6 (2.1)**

Note. Numeric values represent the proportion of discrete-underweighting choices (or CPT-overweighting choices) in experience minus description. Figures in bold are significantly different from zero according to a mixed effects analysis controlling for the random effect of studies and participants.

(*Appendices continue*)

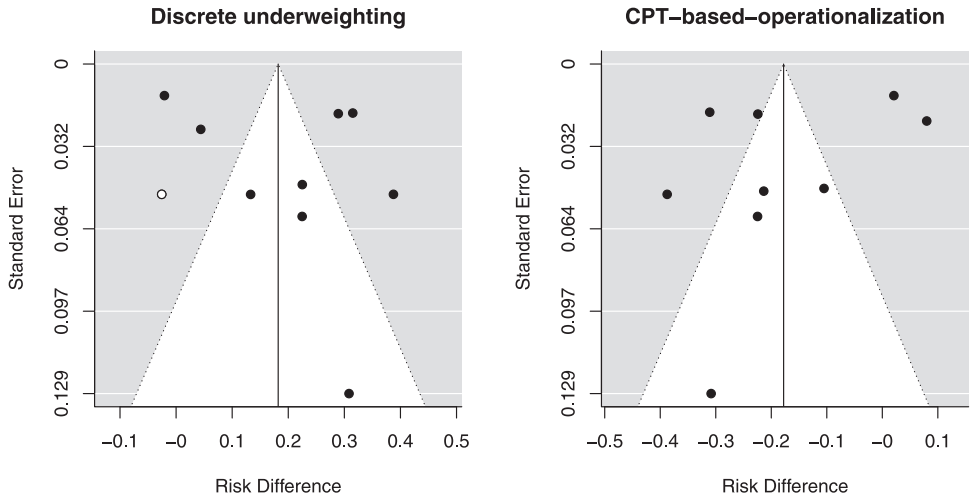

Figure C1. Funnel plots of the description-experience gap in the partial-feedback paradigm (see Figure 2). Plots show the results after imputing missing values (white dots) using the trim and fill method. The vertical line reflects the pooled mean effect size after trim and fill correction. The diagonal lines represent the corresponding 95% confidence intervals.

Appendix D

Recency Results Retrieved from the Literature

Article	Method	Result (as reported; 1st vs. 2nd half)	Interpretation (as reported)	Sampling type	Notes
Hertwig et al. (2004)	Within-option	59% vs. 75%	**Recency**	Autonomous	Same data (VUM reported in
	Value-updating model	φ = .29	**Recency**	Autonomous	Hertwig et al. (2006))
Rakow et al. (2008)	Within-option	66% vs. 76%	**Recency**	Autonomous	—
Hau et al. (2008)	Within-option	58% vs. 60%	No effect	Autonomous	—
Ungemach et al. (2009)	Within-option	65% vs. 59%	No effect	Autonomous	—
	Within-option	42% vs. 48%	No effect	Matched	—
Camilleri and Newell (2009a)	Within-option	56% vs. 61%	No effect	Matched	—
Rakow and Rahim (2010)	Within-option	69% vs. 55%	**Primacy**	Regulated	Adults, Exp. 1
	Within-option	64% vs. 64%	No effect	Regulated	Adolescents, Exp. 2
	Within-option	64% vs. 56%	**Primacy**	Regulated	Adolescents, Exp. 3
	Within-option	58% vs. 57%	No effect	Regulated	Young adolescents, Exp. 3
	Within-option	50% vs. 56%	**Recency**	Regulated	S1, Children
	Within-option	60% vs. 66%	**Recency**	Regulated	S2, Children
Camilleri and Newell (2011a)	Across-option	63% vs. 49%	**Recency**	Regulated	Same data
	Within-option	—	No effect	Regulated	
Camilleri and Newell (2011b)	Within-option	39% vs. 65%	**Recency**	Autonomous	
	Within-option	47% vs. 54%	No effect	Matched	Same data (S1)
	Across-option	—	No effect		
	Within-option	57% vs. 51%	No effect	Matched	Same data (S2)
	Across-option	—	No effect		
Wulff et al. (2014)	Mirror-image	26% vs. 74%	**Recency**	Autonomous	Only discriminating trials
Frey et al. (2015)	Value-updating model	φ = .32	**Recency**	Autonomous	Younger adults
	Value-updating model	φ = .32	**Recency**	Autonomous	Older adults

(*Appendices continue*)

Appendix E

Detailed Analysis of Order Effects

The typical procedure in testing for order effects is to pit recency and primacy against each other and to declare one effect to be the "winner." This approach risks overlooking that both effects can occur, though to different extents. To analyze this risk in detail, we carried out a fine-grained analysis of order effects to detect any co-occurrence of primacy and recency. To this end, we again employed the within-option and across-option methods (see main text; but not the mirror method as it was not appropriate for the current analysis) and compared the predictive accuracy of three (rather than two) parts of the sampling sequence: an *early*, a *middle*, and a *late* part. If an individual's choices were better predicted by the early part, relative to the middle part, a primacy effect would be implied. If an individual's choices were better predicted by the late part, relative to the middle part, a recency effect would be implied. We found no evidence for the co-occurrence of primacy and recency effects. As Figure E1 shows, in both autonomous and matched sampling, the average pattern that emerged was consistent with *increasing* predictive accuracy of the sampled information across both methods. For 78% (within-option method) and 96% (across-option method) of the autonomous sampling data sets, the middle part was more predictive than the early part. For matched sampling, the same held for 75% of sets with both methods (Figure E1). This means that the recency effect for autonomous and matched sampling (see Figure 8) does not "hide" a primacy effect. What about regulated sampling? Does the lack of recency (or primacy; Figure 8) hide the co-occurrence of an equally large primacy (or recency) effect? The predictive accuracies derived from the within- and the across-option methods were 60% and 60% (early part), 62% and 61% (middle part), and 60% and 62% (late part), respectively. These findings seem to suggest that there is no order effect in regulated sampling.

Taken together, we found evidence for recency (but not primacy) in autonomous and matched sampling and no evidence for order effects in regulated sampling. This is consistent with the findings presented in Figure 8. One note of caution, however: In order to conduct this more fine-grained analysis, we had to exclude between 17% and 74% of trials from the various analyses.

(*Appendices continue*)

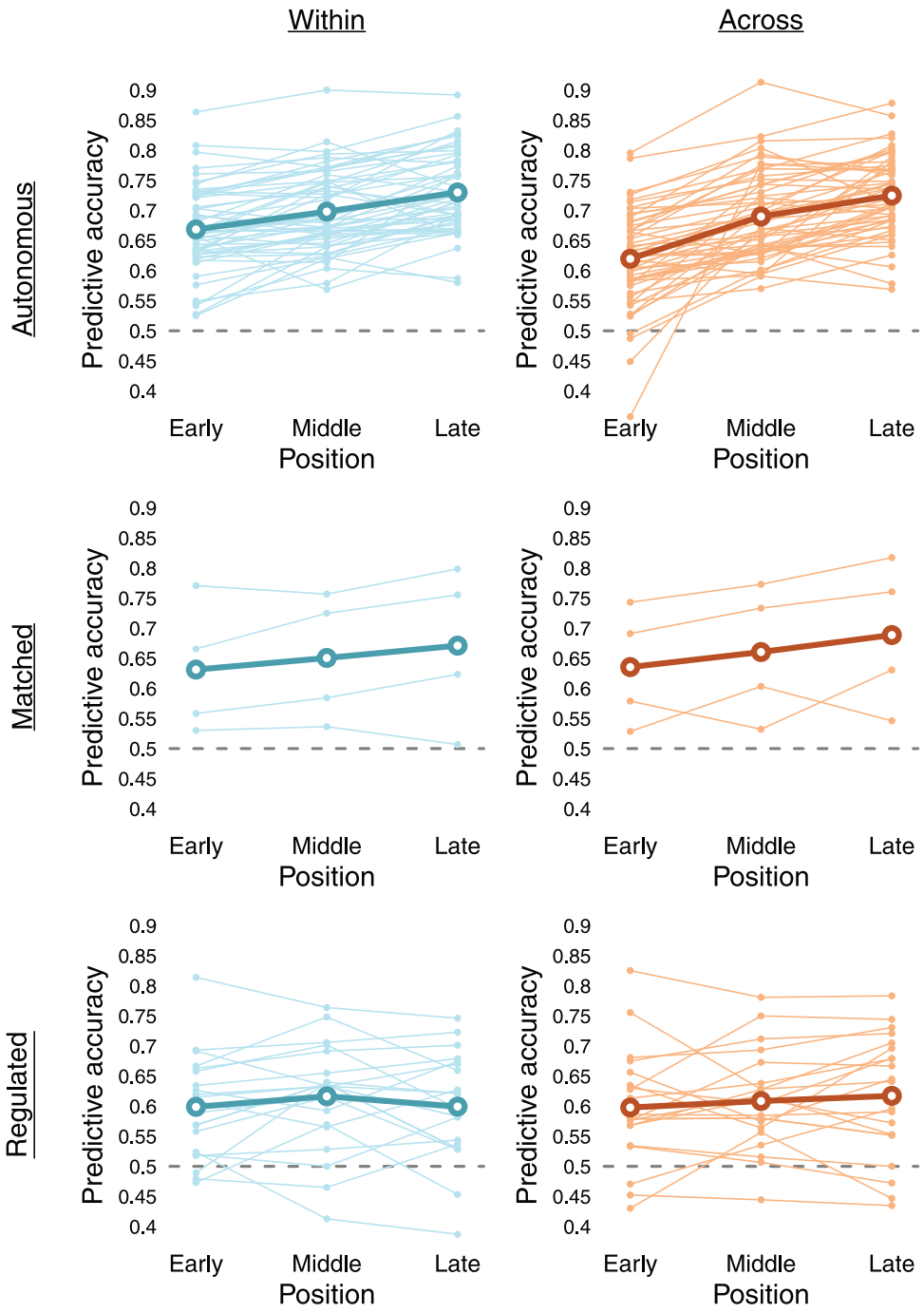

Figure E1. Order effects in the sampling paradigm. The figure shows the capacity of the early, middle, and late part of the sequence to predict final choice in autonomous (upper panel), matched (middle panel), and regulated (lower panel) sampling based on the within-option and across-option method (see main text). The gray lines in the background represent individual data sets and the black line in the foreground their average. The gray dashed line represents random performance. See the online article for the color version of this figure.

(*Appendices continue*)

Appendix F

Recency as a Consequence of Optional Stopping

To demonstrate that optional stopping can lead to recency, we implemented two possible optional stopping strategies. *Stop-when-easy* could be the strategy of a sampler who is very cognizant of the costs of search. She observes the difference between the two options $\Delta = u_A - u_B$ and is more likely to terminate search when Δ is large. Formally, our implementation of stop-when-easy assumes that, at every round i of the sampling process, a person calculates a probability of choosing A:

$$p_{A,i} = [1 + e^{-\phi_i \Delta_i}]^{-1}. \quad \text{(C1)}$$

This probability is a function of $\Delta_i = \sum x_{A,i}/n_{A,i} - \sum x_{B,i}/n_{B,i}$ and of $\phi_i = n_i/\gamma$. The latter implies that p_A, and thus the probability to stop sampling, increases with the size of the sample n_i (relative to a scaling factor γ). The probability of stopping sampling at round i is then calculated as

$$p_{stop,i} = 2 * [\max(p_{A,i}, 1 - p_{A,i}) - .5]. \quad \text{(C2)}$$

Stop-when-complete could be the strategy of a sampler who expects an equal number of outcomes in both options. Therefore, she continues to search whenever fewer outcomes than expected have been experienced for the options. Formally, stop-when-complete terminates sampling at round i according to

$$p_{stop,i|equal} = 1 - (1/n_i)^{\alpha}, \quad \text{(C3)}$$

when the number of experienced outcomes is below k, and according to

$$p_{stop,i|unequal} = 1 - (1/n_i)^{\beta}, \quad \text{(C4)}$$

with $\beta > \alpha > 0$, when the number of experienced outcomes is equal to or above k, with k being randomly drawn from the distribution of problems.

Once sampling has been terminated, both strategies choose option a according to

$$p_{A,n} = [1 + e^{-\Delta_n}]^{-1}, \quad \text{(C5)}$$

and option B with a probability of $p_{B,n} = 1 - p_{A,n}$.

Do these two possible optional stopping strategies produce recency? To find out, we implemented the two strategies in the context of the six decision problems studied in Hertwig et al. (2004) and aggregated results over those parameter values that, according to simulation, produce average sample sizes between 10 and 30 (i.e., $\gamma = [15, 150]$, $\alpha = [0, .15]$, and $\beta = [.32, 1]$). Figure F1 shows the results of 10,000 simulated respondents for each strategy. Both strategies can produce recency (based on the within-option method; see main text), of a similar magnitude and shape as that found in empirical data sets investigating the Hertwig et al. problems (Hau et al., 2008; Hertwig et al., 2004; Wulff & Hertwig, 2017). As shown in Figure F1, both strategies can also produce primacy effects, but the predominant pattern is recency. Moreover, both strategies appear to mimic some, but clearly not all, of the variation found for the six problems evaluated. In other words, this analysis offers an existence proof for the ability of optional stopping strategies to produce order effects—in particular, recency.

Figure F1. Optional stopping and recency. Displayed are the results of the within-option method for two optional stopping strategies (SWE: stop-when-easy; STC: stop-when-complete). Lines show the proportion of choices consistent with primacy (dotted lines) and recency (solid lines) predictions for each of the six problems studied in Hertwig et al. (2004) in the empirical data sets (Hertwig et al., 2004, Study 1 of Hau et al., 2008, and Wulff & Hertwig, 2017; left panel) and according to the stop-when-easy strategy (middle panel) and the stop-when-complete strategy (right panel). See the online article for the color version of this figure.

(*Appendices continue*)

Appendix G

CPT and Certainty in Experience

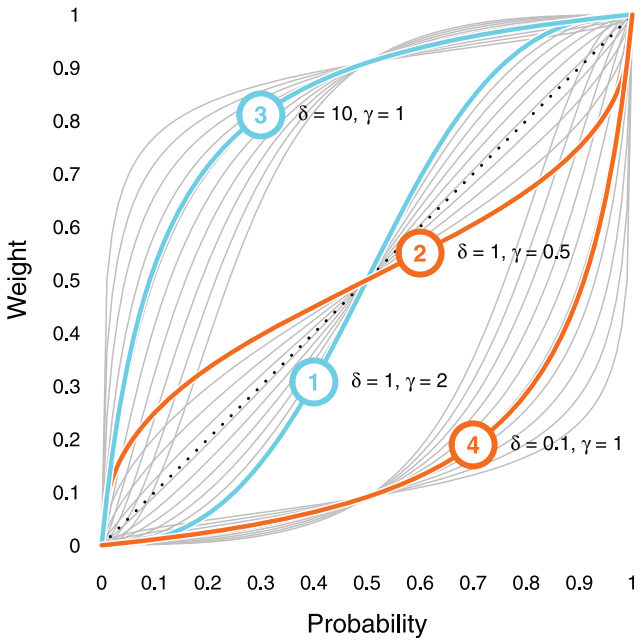

Figure G1. Four shapes of the probability weighting function. The figure shows four weighting functions of the Goldstein and Einhorn (1987) family that illustrate its four qualitative shapes (or subfamilies): S-shaped (1), inverse-S-shaped (2), concave (3), and convex (4). Two of these, shapes 2 and 4, imply a certainty effect, whereas the other two, shapes 1 and 3, do not. Lines in the background illustrate subsets of other parameter combinations. See the online article for the color version of this figure.

Concerning the use of (parametric) weighting functions to measure probability weighting in decisions from experience, there is one fundamental conceptual obstacle that requires better understanding. It concerns the certainty effect. This effect refers to the observation that in lotteries with stated probabilities "people overweight outcomes that are considered certain, relative to outcomes which are merely probable" (Kahneman & Tversky, 1979, p. 265). To capture this effect within the CPT's weighting function, the function's curvature *has to* be convex in the region of large probabilities (see Figure G1, curve 2 and 4). But if so, it would automatically exclude the existence of an S-shaped weighting function (curve 1) indicative of underweighting of small probabilities. Conversely, an inverse S-shaped weighting function (curve 2; as commonly assumed for stated probabilities) implies overweighting of small probabilities and entails a certainty effect. Because of this conceptual dependency, the weighting of low-probability events cannot be measured completely independently of the absence or presence of the certainty effect.

Researchers have argued that experienced-based choice should not give rise to the certainty effect (Barron & Erev, 2003). Unlike with stated probabilities, individuals in experienced-based choice can never be 100% certain that outcomes are truly certain (see Glöckner et al., 2016). If this is the case, then it can only be modeled with an S-shaped weighting function (curve 1; implying underweighting of rare events) or consistent overweighting of the whole range of probabilities (curve 3). If, however, people do reach a level of certainty comparable to that reached under stated probabilities, then it can only be modeled with either an inverse S-shaped weighting function (curve 2, implying overweighting of rare events) or consistent underweighting of the whole range of probabilities (curve 4). This means that the potential existence of a certainty effect in decisions from experience limits the weighting functions available to model (and measure) the behavior to asymmetric subspaces. This makes it impossible to measure the weighting of small probabilities independently of the midrange of the probability scale. For instance, if one assumes the absence of a certainty effect and, simultaneously, overweighting of small probabilities, then this can only be modeled using a weighting function that also assumes an even higher weight for the midrange of the probability scale (curve 3, Figure G1).

Compounding this thorny issue, researchers have recently argued in favor of an inverse S-shaped weighting function in decisions from experience on the basis of a regression-to-the-mean-like process that pulls subjective probabilities toward .5 (Denrell, 2015; Glöckner et al., 2016). When implemented in CPT, however, such inverse S-shaped weighting functions would imply the presence of a certainty effect.

Received May 27, 2016
Revision received April 5, 2017
Accepted May 10, 2017 ■