Value is context dependent: On comparison processes and rank order in choice and judgment

by

Emina Canic

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

October 2016

Table of contents

List of Tables

List of Figures

Acknowledgements

I could not have written this thesis without my supervisor, Neil Stewart. Neil, I think you are awesome. Thank you for putting up with me. I am also grateful to Thomas Hills, who helped me develop research ideas and supported experiments in the field of moral cognition, and Elliot Ludvig, who helped me get published. Both Thomas and Elliot found time to listen and talk to me throughout my time as a PhD student.

Thank you to my parents, who were always there. To my sister and my friends at home, in England, Germany and the US: I am grateful for everything and I appreciate you being and staying in my life.

This research was supported by the ESRC Network for Integrated Behavioural Sciences ES/K002201/1, and the department of Psychology at the University of Warwick.

Declaration

I'm submitting this thesis to the University of Warwick. I have not submitted it anywhere else in any previous application for any other degree. I have completed this thesis independently and I have listed all references.

All chapters are collaborations with my supervisor Neil Stewart. In Chapter 2 and 6, additional colleagues were involved. Chapter 2 was a collaboration with part of the CeDEx lab at the University of Nottingham (Despina Alempaki, Chris Starmer and Fabio Tufano), Tim Mullett here at the University of Warwick and Will Matthews at the University of Cambridge. Chapter 6 emerged out of a collaboration with Marc Scholten from the European University in Lisbon and Daniel Read from Warwick Business School. I am the lead author for all chapters, except chapter 2 where I share lead role with Despina Alempaki. I designed, ran, and analyzed all of the new experiments presented in this thesis, except some of those in Chapter 2 (see Table 2.2 and 2.3).

Abstract

In psychology as well as behavioral economics, it is well established that our choices and judgments are not just a function of the available options, but also of the context surrounding them. Several models have been brought forward to explain these context effects. We use the decision by sampling model (DbS; Stewart, Chater, & Brown, 2006) and investigate possible mechanisms that might lead to the relativity of judgment and choice. Stewart, Reimers and Harris (2015) demonstrated that shapes of utility and probability weighting functions could be manipulated by adjusting the distributions of outcomes and probabilities on offer. Chapter 2 reports a multi-level replication where we find that these effects are robust, but that DbS is unlikely to be the (sole) explanation for its origins. We conclude that problems with revealing utility functions from expected utility fits may be responsible for biasing the shapes of utility functions. Chapter 3 shows that reduced working memory capacity, as manipulated by cognitive load, does not reduce the effects found in Chapter 2. This further points away from a DbS explanation of the above findings. In Chapter 3, we also find that cognitive load has no impact on risk aversion, but find that choice consistency is reduced when working memory capacity is reduced, which also challenges the prominent dual process theories. Still, the question where the differences in preferential functions come from, remains unexplained. Chapter 4 reviews over 20 behavioral as well as neurophysiological studies showing that even if the rank effects are an artefact of the estimation procedure, this does not question the many findings that support a model encompassing a rank-dependent evaluation of alternatives. In Chapter 5, we test this hypothesis in a new design, where a monetary outcome is evaluated in the light of another foregone outcome with a history of other foregone outcomes. In contrast to our hypothesis, we find no evidence for a rank-dependent evaluation here. In Chapter 6, we investigate how comparison processes can lead to the mutable-zero effect. In the mutable-zero effect, participants prefer an outcome entailing a “pay zero” or “lose zero” attribute over an outcome entailing a “receive zero” or “win zero” attribute. We find that the only comparisons that are made with “pay zero” are other payments and that the only comparisons that are made with “receive zero” are other receipts. This process leads to “pay zero” comparing favorably and “receive zero” comparing unfavorably, which in turn leads to “pay zero” options being preferred over “receive zero” options. Given the findings in Chapter 2 and 3 are robust, they point to a general problem with estimating preferential functions using models like expected utility theory or prospect theory. Chapter 3 and 5 were a first attempt at testing a DbS-predicted mechanism: In contrast to our predictions, cognitive load did not decrease context sensitivity. Instead, we observed a slight increase in random choices. Finally, in exploring the mutable zero effect, we added to evidence that comparisons spontaneously happen only within gains, or within losses, but not across gains and losses.

1 Models of Choice

In recent years, some researchers have put forth process based theories in the judgment and decision making literature that stand in opposition to the neoclassical and sometimes even the descriptive theories. Whilst these process-based models have the aspiration to be a more realistic account of judgment and choice, they propose specific mechanisms–like comparison processes or rank ordering– that can be studied more thoroughly. In this thesis, I investigate to what degree comparison processes can drive the construction of choice under risk and the evaluation of prospects or sure outcomes. These insights will eventually help establish the role of comparison processes in human judgment and choice and can be afterwards considered to inform policy, design learning applications or even build machines.

In this chapter I will illustrate the breadth of accounts that have been offered to account for some key choice phenomena, starting at prescriptive models of choice, moving on to descriptive models, and finally to process models. Even for a phenomenon as simple as risk aversion, I will review how some accounts appeal to the notion of diminishing marginal utility, some accounts appeal to probability weighting, and some accounts appeal to heuristic processes. Table 1.1 summarizes each phenomenon and the corresponding explanation in each account.

Early models of judgment and choice under risk are based on normative principles, and describe what the rational homo economicus should choose. Expected value (EV) is probably one of the most prominent models of rational decision making and the basis of many subsequent prescriptive as well as descriptive models of choice. Using solely the EV of an option as a criterion to make a choice would mean choosing the option with the highest expected outcome–that is, the option which, if played out multiple times, would give the highest average payoff. There are several reasons why EV is a sensible strategy when making decisions under risk. The aggregate is of essence here: If the same decisions are made repeatedly, if many different kind of decisions are made, or if different people make a

decision in the EV manner, it will maximize EV when the events are all added up. The crux here is that this argument is only really true, if people live infinitely, if there are infinite kinds of decisions and if there is an infinite number of people. However, none of these is the case (Baron, 2007). Why the EV should not be a sufficient criterion for the chooser and why it is not a description of people's choices, is nicely demonstrated in the St. Petersburg paradox: If hypothetically one would offer somebody a lottery to win £2 pounds if a coin lands on tails on the first toss, £4 if it lands on tails on the second toss, £8 if it lands on the third toss etc., according to the EV criterion, everybody should be willing to pay an infinite amount of money to play the gamble, because the EV is infinite. However, to people this lottery is worth only a small amount of money. The fact that people would not pay a lot to play this lottery shows that EV alone is an insufficient criterion for describing choice under risk.

Daniel Bernoulli (1738) offered a resolution for the problem by introducing a (logarithmic) utility function, which transforms EV into expected utility (EU). The function takes into account the money a chooser already has when making a decision and assumes diminishing marginal utility. Diminishing marginal utility means that the utility of every extra unit of wealth is always positive, but is smaller than the utility of the previous unit. This assumption solves the paradox, because while $\sum_i^\infty (\frac{1}{2}^i) 2^i$ is infinite, $\sum_i^\infty (\frac{1}{2}^i) log 2^i$ is finite at 1.39. In addition to resolving the paradox, assuming diminishing marginal utility also nicely captures risk aversion: When offered £100 for sure or £200 with a 50% chance and otherwise nothing, many people will choose not to gamble and take the safe option. This choice is sensible if the utility of £100 is more than half the utility of £200. Although Bernoulli's resolved this paradox already in the 18th century, his hypothesis only became popular when von Neumann and Morgenstern (1947) showed that the expected utility hypothesis follows from simple assumptions or *axioms* about people's preferences (Starmer, 2000, for a review). An individual's utility function can always be estimated when the following axioms are met: 1) when a person has well defined preferences, i.e. the person either likes A over B, B over A

or is indifferent, 2) when a person's choices are transitive, i.e. if she likes A better than B, and B better than C, then she should like A better than C, 3) when stochastic dominance is met, i.e. when she chooses A over B, if A is better in at least one aspect and equally or at least as good in other aspects and finally 4) when her preferences are independent of the method they were elicited or of the way the options were described. These axioms form the basis of normative models.

Given a person does not violate any of these axioms, the person is said to reveal the utilities of the options open to them through making choices that maximize her utility. Using the person's choice behavior, all options can be assigned a value under the EU model. For example, von Neumann and Morgenstern showed how the utilities of three options could be determined by repeatedly offering people a choice between one option, offering a probability mixture of the best and the worst vs. another option offering the middle-ranking option for sure. By adjusting the probability of the best in the mixture up when people choose the sure option, and adjusting it down when people choose the mixture, one can titrate in upon the probability that makes people indifferent. This probability value is then the utility of the middle-ranking option, with the best option arbitrarily being assigned utility 1 and the worst having utility 0. In this way, every option available can be assigned a real number on the cardinal utility scale.

Table 1.1

Choice phenomena and corresponding explanations of different accounts.

	EUT	**(C)PT**	**Regret Theory**	**TAX**	**Priority heuristic**	**Decision by Sampling**
Allais Paradox (common ratio)	–	S-shaped probability weighting	Disproportionately more regret if difference between states of the world is large as opposed to small	Transfer of attention to smaller outcome	If minimum gains differ by 10%, choose higher minimum gain	A combination of the distributions of amounts and probabilities in the world together with a change in the distributions across the scaled up and scaled down choices
Risk Aversion	Diminishing marginal utility	Diminishing marginal value	Diminishing marginal choiceless utility	Transfer of attention to smaller outcome	If minimum gains differ by 10%, choose higher minimum gain	Rank position of attribute values with positively skewed distribution for gains
Loss Aversion	–	Transformation of losses, which loom larger than gains	Convex regret function	Transfer of attention to losses	If minimum losses differ by 10%, choose lower minimum loss	Rank position of attribute values with negatively skewed distribution for losses
Isolation Effect	–	Reference point/status quo matters, not final wealth	Consequences are state-contingent, leading to different initial situations despite identical final wealth	Transfer of attention to smaller outcome	Choose the lower of the two minimum losses and the higher of the two minimum gains if difference exceeds 10%	Separate samples for gains and losses
Violation of Gain-Loss Separability	–	–	–	More weight on losses in mixed gamble than in losses-only gamble	–	–

Yet, people often violate the rationality principles and do not choose the way they ought to (Birnbaum, 2008, for a recent review). Hence EUT does not provide a very accurate description of people's choice behavior either. In the remainder of this chapter, I will outline some of the key departures from EUT, and will start with a classic example, the Allais paradox. What the St. Paradox is for EV is the Allais paradox (Allais, 1953) for EUT: Consider the following choice options:

Option A: win 1Million for sure,

or

Option B: win 5 Million with a 10% chance, or win 1 Million with an 89% chance and otherwise nothing.

People more often choose the safe option A, perhaps avoiding option B because they are scared of the 1% chance of receiving nothing. Now consider these modified options:

Option C: win 1 Million with an 11% chance and otherwise nothing,

or

Option D: win 5 Million with a 10% chance and otherwise nothing.

Now people more often choose the riskier Option D–as if they discount the difference between the 10% and the 11% chance of winning. This preferential pattern describes the common consequence effect. Removing an 89% chance of £1M from both options A and B gives options C and D. Thus, to be consistent with EU, people must choose either options A and C or B and D, because otherwise they violate independence from irrelevant alternatives.

The common ratio effect–also first demonstrated by Allais–deals with an EUT-inconsistency that considers probabilities instead of outcomes:

Option A: win £3000,

or

Option B: win £4000 with an 80% chance, otherwise nothing.

Again most people choose the safe Option A. On the other hand, when asked to choose between these modified two options,

Option C: win £3000 with a 25% chance, otherwise nothing,

or

Option D: win £4000 with a 20% chance, otherwise nothing,

people prefer the riskier Option D. Here, people are sensitive to the probability differences between Option A and B, but discount the difference between Option C and D, although the ratio of the expected utilities for A and B is identical to the ratio for C and D.

Not only do people violate the independence axiom, but the way options are described also influences people's choice behavior in a non-EUT way. People reverse their preferences if options are presented as losses instead of gains. The Asian disease problem (Tversky & Kahneman, 1984) illustrates this framing effect by asking people to choose between two scenarios described in different ways: The relatively safe option describes that program A will either save 200 out of 600 people or it describes that it will kill 400 out of 600 people. The relatively risky option describes that program B will safe 600 people with a 33.33% chance, otherwise nobody will be saved or it describes that nobody will die with a 33.33% chance, otherwise everybody will die. The first versions of the programs are framed as gains, by using the word "save". The second versions are framed as losses by using the words "kill" or "die". Note that all options have an identical EV and that the outcomes are objectively the same whether framed as gains or losses. However, people more often choose the safe option A when using a gain frame (scenarios using the word save) and prefer the

relatively risky option B when using a loss frame (scenarios using the word die). This clearly violates description invariance.

How can we capture these violations of von Neumann and Morgenstern's (1947) compiled axioms, which also form the basis of EUT? Throughout the upcoming years there were different kinds of modifications. Three modifications in particular were gathered together within prospect theory (Kahneman & Tversky, 1979), a model that has become extremely influential. One type of modification was to introduce probability weighting to transform objective probabilities and integrate them with the corresponding outcomes. One of the earliest models, is Edwards subjective expected utility theory (SEU; 1955, 1962). EUT is a special case of SEU, if the subjective weights in SEU are set to their objective probabilities. As an aside, further developments led to Quiggin's rank-dependent expected utility theory (RDUT; 1982). The main difference between the two is that whilst SEU assigns the same weight to a specific probability independently of the magnitude of the outcome, in RDUT the weight assigned to a specific probability is dependent on the magnitude of the outcome, specifically, how good this outcome is in relation to other possible outcomes within the same prospect. Further modifications added psychological insights into how people think about money. Kahneman and Tversky suggested people think about changes in wealth, not final wealth states. Thus people consider the gains and losses on offer, and fail to integrate them into final wealth states. They also considered that people display diminishing sensitivity from the reference points and from extremes, motivating the shapes of their value and weighting functions. Finally, Kahneman and Tversky introduced the notion of loss aversion, where losses loom larger than gains. Together these modifications are the basis of prospect theory (PT, Kahneman & Tversky, 1979).

How does PT deal with the common-ratio and the framing effect presented above? PT describes a choice under risk in terms of an inverse S-shaped weighting function, and a value function with zero as a reference point. Often this zero is taken to be current wealth. The prospects are evaluated as losses and gains relative to the reference point, instead of the integration of these states into total wealth as in EUT. The weighting function allows objective probabilities to be transformed into subjective weights by underweighting high probabilities and overweighting low probabilities. This means that people are most sensitive at the extremes, i.e. they distinguish 5% from 10%, but are insensitive in the middle, i.e. 45% and 50% are treated similarly. Probability weighting is how PT accounts for the common-ratio effect: Because people are very sensitive to probability differences at the extremes, 80% seems very different from 100% and thus makes the safe £3000 much more attractive than the risky £4000. On the other hand, when presented two risky options, where the chances of winning are 20% and 25%, which appear very similar to each other in a PT-account, the prospect of winning £4000 seems now much more appealing than a chance to win £3000.

The value function transforms absolute amounts into subjective values, by measuring the change from the relative reference point, the status quo. The value function is kinked at the reference point and all amounts above it are translated into subjective utilities using a concave function, so that amounts further away from the reference point are treated similarly. All amounts below the status quo are translated into subjective utilities using a steeper, convex function (see Figure 1.1). The loss function is convex only because it is upside down. It is still the case that a loss of £100 vs £200 is a larger difference than £1100 vs £1200. A concave value function is how risk aversion was accounted for in the traditional EUT-manner. The new

property here is that the reference point serves the evaluation of prospects as gains and losses.

PT's value function nicely describes risk seeking behavior in the loss domain and risk averse behavior in the gain domain, as we saw in the Asian disease problem. When the programs are presented using a gain frame, people preferred the program that saves 200 people for sure over the program that saves 600 people with a 33.33% chance. If 200 lives saved for sure are evaluated using a concave function, they are worth more than 600 lives saved maybe. However, when the programs are presented using a loss frame, people preferred the program, which kills 600 people with a 66.66% chance to the program that kills 400 people for sure. Losing a minimum number of people is so bad that one is willing to take the risk if one can avoid a loss altogether.

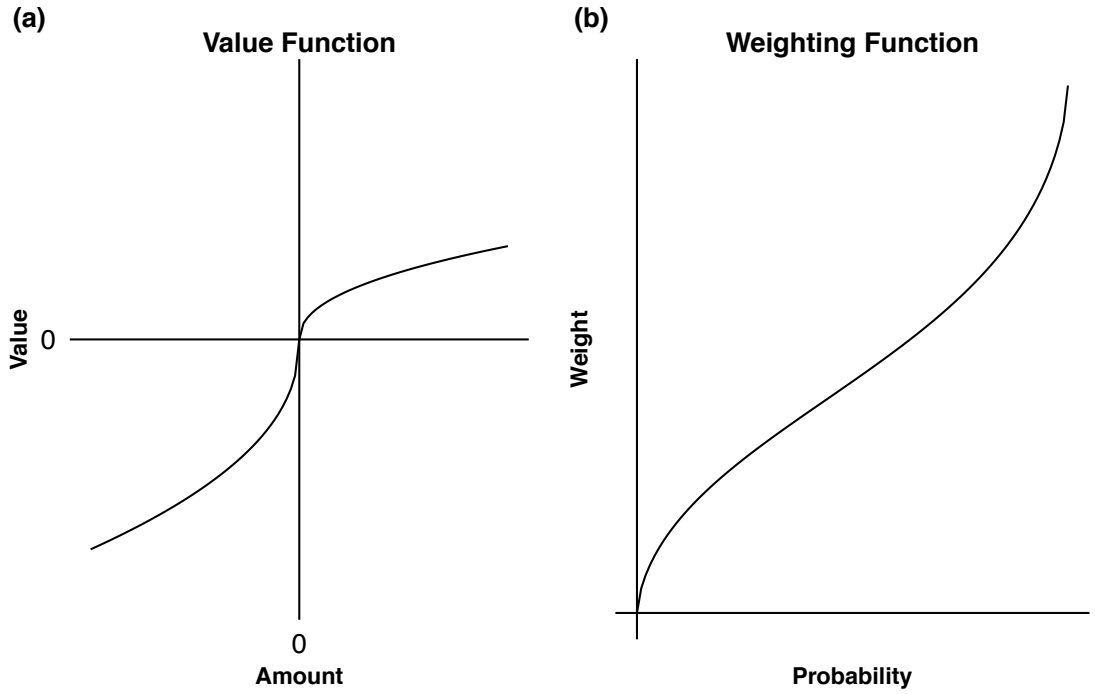

Figure 1.1. Prospect theory's value and probability weighting functions.

The fact that people's perspective matters and not final wealth is evident in the isolation effect: If a person is given $100 and is then asked to choose between

winning another $50 for sure or another $100 with a 50% chance, the person is more likely to choose the safe option. On the other hand, if a person is given $200 and is then asked to choose between losing $50 for sure or losing £100 with a 50% chance, the person is more likely to choose the risky option. If people integrated their initial gift of $100 into the choice and considered their final wealth, the two choices would be the same. Gains are treated differently from losses, because "losses loom larger than gains". Because the impact of a loss is stronger than the impact of a gain, people are described to be loss averse; they prefer a sure gain in the light of a possible loss, even if the expected value is higher than the sure option can offer.

PT without editing rules ("stripped PT" in Birnbaum, 2008) can violate stochastic dominance. In stripped PT the value of a prospect where one can win £100 with 50% chance and otherwise £100 is smaller than the value of a sure £99. This is because of probability weighting: The weight for 50% is much smaller than an objective 50% so that when integrated into the £100 prizes the value adds up to less than £99. In order to avoid that, Kahneman and Tversky (1979) invented some editing rules. The prospects are edited through a) combination of probabilities if they are associated with the same outcome (this would be the case in the above example), b) separation of outcomes that are riskless, or c) if two prospects share an outcome, they cancel each other out, d) probabilities and outcomes are simplified by rounding and finally, e) when an alternative is dominated, it needs to be discarded. PT's editing rules become obsolete when probabilities are transformed into cumulative decision weights, like they are in RDUT (Quiggin, 1982). Tversky and Kahneman (1992) exchanged their former weighting function and editing rules with a cumulative probability weighting function and developed a more general model, cumulative prospect theory (CPT). CPT, as opposed to PT, is applicable for gambles with more

than two non-zero outcomes and in CPT different weights can be assigned to identical probabilities depending on whether an outcome was perceived as a loss or a gain. CPT does this by assuming people work with the probability of doing at least as well as £X instead of the probability of receiving £X–that is cumulative probabilities over a set of events are psychologically elemental, instead of the probabilities of individual events. With this new feature CPT is able to explain risk seeking and risk averse behavior or the four-fold pattern within one person by allowing different weighting functions for losses and gains (Tversky & Kahneman, 1992). The four-fold pattern describes risk seeking behavior when dealing with large prizes and small probabilities, and risk averse behavior when dealing with small prizes and low probabilities considering low stakes. In the loss domain, the same participant will be risk seeking when dealing with low stakes and will be risk averse when dealing with high losses. A person is risk-averse if she chooses the safe option over a risky option with an equal expected value. A person is risk-seeking if she chooses the relatively risky option under the same circumstances. If the same person chooses between a safe and a risky option in the loss domain and is risk-seeking and then is asked to choose between a safe and a risky option and plays for the same amount of money in the gain domain and is risk-averse, the person displays the reflection effect.

A little after PT was proposed, Loomes and Sudgen (1982) developed regret theory to make predictions in line with the Allais paradoxes, risk aversion and loss aversion. In regret theory, there is a number of states of the world, which each occur with an objective probability that is known to the chooser, or stands for the chooser's belief about the occurrence of each possible state. Given one action, there is one consequence for each state of the world. One assumption that Loomes and Sudgen make is that different people have different choiceless utility functions, just like in

expected utility. The choiceless utility function assigns a utility to a consequence that the chooser has not chosen herself. The consequence is either inflicted or awarded by somebody else. The utility of choosing an option A, given a particular state of the world, will be the sum of its choiceless utility and the difference between the joy of choosing A and not B, and the regret of choosing A and not B, all assuming the same state of the world, and considering that regret is convex in outcome difference. This means that large differences in outcomes in transformed into huge amounts of regret. Taken together, this means that the evaluation of option A is affected by the presence of option B. Thus, the evaluation is based on the comparison between the options available, given a specific state of the world.

Note that the (C)PT and regret explanations of the common ratio effect are quite different. (C)PT explains the common ratio effect by means of an inverse S-shaped probability weighting function. In contrast to that, regret theory predicts this effect by assuming that there is disproportionately more regret if the difference between what is and what might have been is large, as opposed to if the difference between what is and what might have been is small. In the scaled up version of the common ratio effect, the largest possible regret is caused when you choose B and get £0, when choosing A would have delivered a sure £3,000. But in the scaled down version, it is possible to choose C and receive £0 when choosing D could have offered £4,000. This difference is larger, and because of the convexity of the regret function, the anticipated regret of missing out on £4000 dominates the decision. In summary, scaling down the problems introduces new and particularly regretful comparisons which reverse the preference. In this framework, anticipated regret is also the main driver for the reflection effect: The relatively safe option as compared to a riskier option is more appealing in the gain domain, because potentially losing out on a

possible gain can lead to regret, which the decision maker wants to avoid, whereas the relatively risky option is more appealing in the loss domain, because of the higher possibility to avoid a loss.

The models presented so far are all based on the standard normative models that were later modified to account for empirical data. (C)PT and regret theory used psychological insight to adapt EUT. Other modifications like RDUT only aimed at explaining empirical phenomena and were not concerned with the influence of the human factor in building their models. There is another class of descriptive models, which approaches the paradoxes and the choice reversals from a psychological perspective. In these configural weight models, the weight of an outcome in one prospect depends on other outcomes in the same prospect. Configural weight models are somewhat similar to rank-dependent utility models (Birnbaum & Navarette, 1998), but also share similarities with rank-dependent weighting models like CPT. In configural weight models utilities of options are assumed to be weighted averages of all possible outcomes per option. The weights depend not only on the outcomes themselves, but also on the ranks, the probabilities of those outcomes and the point of view of the judge, e.g., if the judge is a buyer or a seller, or how risk-averse the person is. TAX, the "transfer of attention and exchange" model (Birnbaum & McIntosh, 1996; Birnbaum & Stegner, 1979), is a special case of one of these models, which predicts the paradoxes described above, the preference reversals and the four-fold pattern. (In addition, TAX predicts many additional phenomena that CPT cannot account for, see Birnbaum 2008 for a review.) In TAX, a weight is the attention a decision maker pays to each outcome. An outcome that is more probable should also get more attention, but depending on somebody's risk attitudes, the attention gets redistributed to other smaller outcomes. For example, consider a choice between a

50/50 chance for £100 otherwise £0, or £50 for sure. We know people will tend to be risk averse, and select the sure £50. In EU and prospect theory this risk averse behavior is explained using a concave utility function where the utility of £50 is more than the average of the utilities of £0 and £100. But in TAX, attention moves from the £100 outcome to the £0 outcome, and this movement of attention from the higher to the lower outcome reduces the attention weighted sum of the values, making this option less appealing. The same logic is used to explain why people behave as if they are loss averse: If people are unwilling to accept a bet where they could either win £200 or lose £200, one only has to assume that people place more attention to "lose £200" than to "win £200" to give this bet a net negative evaluation and reject it. It is not necessary to transform money into utility and have losses loom larger than gains, but just to assume that it is the transfer of attention–weighting–from the better outcomes to the worse outcomes–that causes risk-averse or loss-averse behavior (Birnbaum & Navarette, 1998; Birnbaum, 2008). Transfer of attention is also how TAX predicts the common ratio effect: When choosing between a safe gamble without the possibility of a £0 outcome, and a relatively risky gamble with the possibility of a £0 outcome, the weight within the risky gamble shifts from the possible amount of £4000 to £0, making the riskier gamble less attractive than the safe gamble with the £3000 win and prompting more risk averse choices. When choosing between two risky options that both involve the possibility of a £0 outcome, and evaluating each option separately within one option, attention is transferred from a possible gain to £0. Now that both options have a drawback, the riskier option becomes more appealing because it offers a higher win, prompting more risk-seeking choices.

There are several studies demonstrating inconsistencies with neoclassical rank-dependent theories–CPT is among those theories– that simple configural weight models can account for (see Birnbaum, 2004, and Birnbaum, 2008 for the most recent paradoxes). Gain-loss separability, which empirically is found to be violated, but is predicted by the neoclassical theories, is one of these examples. Wu and Markle (2008) asked participants to choose between gamble A+ and gamble B+ in the gain domain and gamble A- and gamble B- in the loss domain. They found that although a majority of people preferred gamble B+ to A+ and gamble B- to A-, when gamble A+ was mixed with A- and gamble B+ was mixed with B-, people suddenly preferred A-mixed to B-mixed–that is in the gain domain participants preferred Option B+ when choosing between

Option A+: win £2000 with a 25% chance or win £800 with another 25% chance and otherwise nothing, or

Option B+: win £1600 with a 25% chance or win £1200 with another 25% chance and otherwise nothing.

They also prefer B- when choosing between

Option A-: lose £800 with a 25% chance or lose £1000 with a 25% chance and otherwise nothing, or

Option B-: lose £200 with a 25% chance or lose £1600 with a 25% chance and otherwise nothing.

Expectedly, people are risk averse in the gain domain and risk seeking in the loss domain. However, if the gain gambles are combined with the loss gambles, people suddenly reverse their previous preferences by choosing option A out of

Option A: win £2000 with a 25% chance or win £800 with another 25% chance, or lose £800 with a 25% chance or lose £1000 with a 25% chance,

and

Option B: win £1600 with a 25% chance or win £1200 with another 25% chance or lose £200 with a 25% chance or lose £1600 with a 25% chance (see what gambles Option A and B are composed of in Table 1.2).

By preferring B+ and B-, but also A, people violate gain-loss separability (see also Birnbaum & Bahra, 2007). Thus, whilst neoclassical theories qualify to explain findings in single-domain gambles, they fail in mixed gambles. At the same time, other models–like TAX–can predict and can account for the violation of gain-loss separability or stochastic dominance (see Birnbaum, 2008). If we judge theories of choice against empirical evidence and we find that some very prominent theories–in economics and psychology–can be refuted based not only on violations of normative principles, but also on assumptions made that are not compatible with people's choices, should we still use those models in appropriate contexts?

Table 1.2

Options A and B result from merging Option A+ with A- and Option B+ with B- respectively. The probabilities to win or lose are 25%. In the one-domain options, the gambles offer a 50% chance of nothing.

Option A	Nonzero Option A+ gains		Nonzero Option A- losses	
	+ £2000	+ £800	– £800	– £1000
Option B	Nonzero Option B+ gains		Nonzero Option B- losses	
	+ £1600	+ £1200	– £200	– £1600

Up to this point the presented models were silent about the information integration and the choice process at hand when faced with a decision problem. There is another class of models, which does propose a specific cognitive process for different problems ranging from inference to preference. For inference tasks Simon

(1955) proposed that people do not or cannot optimize, but use simple choice rules–heuristics–to make decisions; because there are constraints to human cognition, which make it difficult to adhere to normative principles. To reduce the requirements for the decision maker, the main feature of the proposed heuristics is the neglect or disregard of some information. The heuristics differ in terms of which information is ignored. Several researchers argued that by ignoring information when making a decision, people are able to make even better, or more robust decisions than if they took all available information into account (Gigerenzer & Todd, 1999).

The decision making literature has benefitted from the formulation of process models. Already back in the 70s researchers have encouraged process-oriented models (Payne, 1976; Payne, Bettman, & Johnson, 1993). They argued that collecting process data will guide us to more accurate models of choice and will help make much quicker progress in this endeavor (Johnson, Schulte-Mecklenbeck, and Willemsen, 2008). The priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006) will represent this class here, although as process model as well as descriptive model it has been convincingly refuted (see Birnbaum, 2010, or Reiger & Wang, 2008, and also other references below).

Heuristics are process models: they describe step-by-step how people supposedly make decisions and are perceived as decision making tools, where each heuristic can be selected depending on the choice problem and where different people might choose different heuristics for one and the same task. For tasks involving choices between two options, there are two classes of heuristics that could be considered: Tallying and lexicographic heuristics. Tallying works by looking through all attributes of both options in any order and whichever option has more satisfying attributes, gets chosen. For the lexicographic heuristic, different attributes of both

options are first ranked and whichever option has the highest value on the most important attribute, gets chosen (Tversky 1969).

To account for some of the above mentioned violations of EU theory, like the Allais paradox, the reflection effect, the fourfold pattern or intransitive choice preferences, Brandstätter, Gigerenzer, and Hertwig (2006) have proposed the priority heuristic (PH). Using PH, people consider the attributes of a gamble in the following order: minimum gain, the probability of the minimum gain and lastly the maximum gain. When the minimum gains differ by at least 10% of the maximum gain, people stop the evaluation process and choose the option with the higher minimum gain. If the minimum gains are too similar, then people compare the corresponding probabilities and choose the gamble with the higher probability of the minimum gain, if it exceeds the other by 10%. If this too is not the case, they choose the gamble with the higher maximum gain. The heuristic works the same way for the loss domain, where the gamble with the lower losses and higher probability of the minimum losses are chosen.

To predict the fourfold pattern, consider the following two options. Option A offers a 95% chance of £100 otherwise nothing and option B offers £95 for sure. First, the minimum gains are compared; £0 from option A and £95 from option B. £95 (the higher minimum gain) is higher than 10% of £100 (the maximum gain) so that now according to the heuristic, people would choose the sure option B, displaying risk aversion. The priority heuristic would explain risk seeking choices in the exact same way if the above gambles would involve losses. This is also how the common ratio effect works in PH: When choosing between two gambles,

Option A: 80% chance of winning £4000, otherwise nothing,

or

Option B: 100% chance of winning £3000,

PH predicts that people choose the safe gamble B, because the minimum gain of B (£3000) is by more than 10% of the maximum gain (10% of £4000 = £400) higher than £0. If these gambles are transformed into

Option C: 20% chance of winning £4000, otherwise nothing,

or

Option D: 25% chance of winning £3000, otherwise nothing,

PH predicts that people choose the riskier option C. Because the minimum gains are £0 in both options, and the probabilities of the minimum gains do not differ by at least 10%, the option with the higher maximum gain will be chosen.

The priority heuristic has been criticized on several levels. There are several studies showing that people do not deal with one cue at a time (Birnbaum, 2008; Hilbig, 2008), that choice patterns and reaction times are more supportive of strategies that integrate probabilities with outcomes (Ayal & Hochman, 2009; Glöckner & Betsch, 2008), and that probabilities, which are ignored by PH systematically influence people's choices (Fiedler, 2010). Birnbaum (2008) also showed that PH failed to perform on a significant proportion of the paradoxes that he himself brought forward and was always outperformed by TAX. Hilbig (2010) pointed to a general problem considering the validity of heuristics: Even if the predictions of the simple choice strategies come true, it does not mean that those were the rules that participants used to make the decision with. Johnson, Schulte-Mecklenbeck, and Willemsen, (2008) take the same line and strongly argue for the collection of process data when investigating process models; it is the only way to fully evaluate the PH. According to the process data these authors collected, the way

participants approach the choice problems varies across individuals and across the different types of gambles and cannot easily be built into the PH.

The first models I presented are built on normative principles. In their development, those principles have been relaxed to account for more empirical data whilst abiding to as many normative rules as possible and dropping as many as necessary; all in the economic tradition. The TAX model, while not defined as a process model by its author, but as an as-if model, shifts the interpretation of options away from being perceived as prospects to trees with branches and proposes that the shift of attention to different possible outcomes is what leads to violations of normative principles and the observed choice behavior. What TAX has in common with the earlier models like PT, regret theory, and EU is the assumption that amounts are transformed into utility, or value by stable and well-defined functions. However, there is an increasing range of research (including the experiments presented later on) challenging this assumption of the stability of value functions. These challenges are at least partly met by decision by sampling (DbS, Stewart, Chater, & Brown, 2006), a process model based on simple psychological principles that combined lead to a context-dependent evaluation of options.

The formulation and integration of the information process during choice often resembles a normative economic account rather than a model based on psychological principles. Models within psychology itself (although not in the adaptive decision maker literature) often even assume that value is computed (Vlaev, Chater, Stewart, & Brown, 2011), for example by multiplicative integration of transformed attributes. Those transformations can depend on other available attributes in the option (like in TAX or, to some extent, CPT), or solely on a judge's risk attitude (like in PT).

Opposed to this account, in other models, attributes are evaluated relative to other attributes so that the value of an option stands in relation to other available options in the choice set. In this account, value is not computed by using an internal map that transforms probabilities into weights and amounts into subjective value. Instead, comparisons between alternatives contribute to the evidence for one alternative over the other, so that the value of an option on its own is not ever even calculated or used at any point during a choice (Vlaev, Chater, Stewart, & Brown, 2011).

In these models value is dependent on the comparison process, either with the immediate context or with attributes from long-term memory. All experiments presented in this thesis are constructed to test whether, when, and to what degree comparison processes influence choices between two gambles, ratings of gambles or simply monetary outcomes. Our hypotheses are based on predictions from decision by sampling (DbS, Stewart, Chater, & Brown, 2006). In DbS, the rank order of an attribute value within a decision sample, determines the subjective value of an option, without people ever calculating or constructing the subjective value of an option. Instead, DbS is formulated as a process model, where in order to evaluate an option, people compare it with other options on offer and keep track of the number of times the current option has won across a series of pairwise comparisons. Thereby people behave *as if* they use the subjective value when they make a decision. For example, in a sample of £3, £5, £6, £10, £13, the subjective value of £13 is higher than in the set of £6, £13, £16, £19, £23, because £13 wins all binary comparisons in the first set and receives the value 1, whereas it only wins one comparison out of four possible comparisons in the second set and receives the value 1/4 there. The higher the attributes of options rank, the more likely they accumulate sufficient evidence to be

chosen among other available options. The rank hypothesis suggests that the subjective value of any given option is determined by the distribution of outcomes available in the choice set and/or in long-term memory.

Stewart, Chater, and Brown (2006) have shown that assuming pairwise comparisons, frequency accumulation, and sampling from memory, combined with the distributions of credits and debits in the world, is all you need to produce the shape of a typical psychoeconomic function–concave value function for credits, convex value function for debits. The real world distributions of amounts serve as approximation of what amounts are in people's long-term memory. That is, to the extent that memory is adapted to and reflects the structure of the environment (Anderson & Schooler, 1991), properties of the distributions of attribute values in the environment could be a useful proxy for the properties of the distributions of attribute values in people's memories. What Stewart et al. found in credits and debits to and from bank accounts, which they used as a proxy for the distributions of gains and losses in people's memories, is that there are more small than large credits (i.e., a positively skewed distribution), and more small than large debits. The positively skewed distributions mean that for a fixed size increase, the effect will be larger if the increase is lower in the distribution than if it is higher in the distribution. For example, because there are more small gains than large gains, an increase of £100 from £0 to £100 overtakes more of the sample than an increase of £100 from £1000 to £1100. This is because there are many gains that lie in between $0 and £100, but few gains lie in between £1000 and £1100. The positively skewed distribution of gains in the real world combined with the decision-by-sampling principles gives a reasonable account for the diminishing marginal utility that is assumed in neoclassical theory and empirically observed. The logic works the same way for losses.

Recently, Stewart, Reimers, and Harris (2015) have experimentally tested the following DbS prediction: when participants encounter attribute values (e.g., amounts, probabilities or times) that are positively skewed, with many smaller attribute values at the low end of the range, the estimated psychoeconomic functions should come out more concave than when participants encounter attribute values that are negatively skewed, with many larger attribute values at the high end of the range. The psychoeconomic functions should also come out more concave than when attribute values are uniformly distributed. According to the rank principle, the utility of an attribute increases quickest where the distribution is densest. Finding the predicted functional forms, means that just by changing what distributions of attributes people have to deal with, is what will make them prefer more or less risky options. The authors asked participants to choose between pairs of gambles, whose outcome distributions were either positively skewed or the participants chose between pairs, whose outcome distribution was negatively skewed. As predicted, the recovered utility functions had a concave form when the distribution of outcomes was positively skewed, whereas it was convex when the distribution of outcomes was negatively skewed. This means that the subjective value of a presented attribute was higher when it was presented among a positively skewed distribution where it occupied a relatively high rank position than when it was presented among a negatively skewed distribution where it occupied a relatively lower rank position. The analogous procedure was performed for probabilities and also times separately, reaching the same conclusion.

This means that people might not be inherently risk averse or loss averse, but that they just appear as if they were, because of the world we live in. With this logic, Walasek & Stewart (2015) tested a possible cause for loss aversion. Could it be that

because of the way losses and gains are distributed in the real world, people behave as if they were loss averse and only therefore displayed loss aversion?

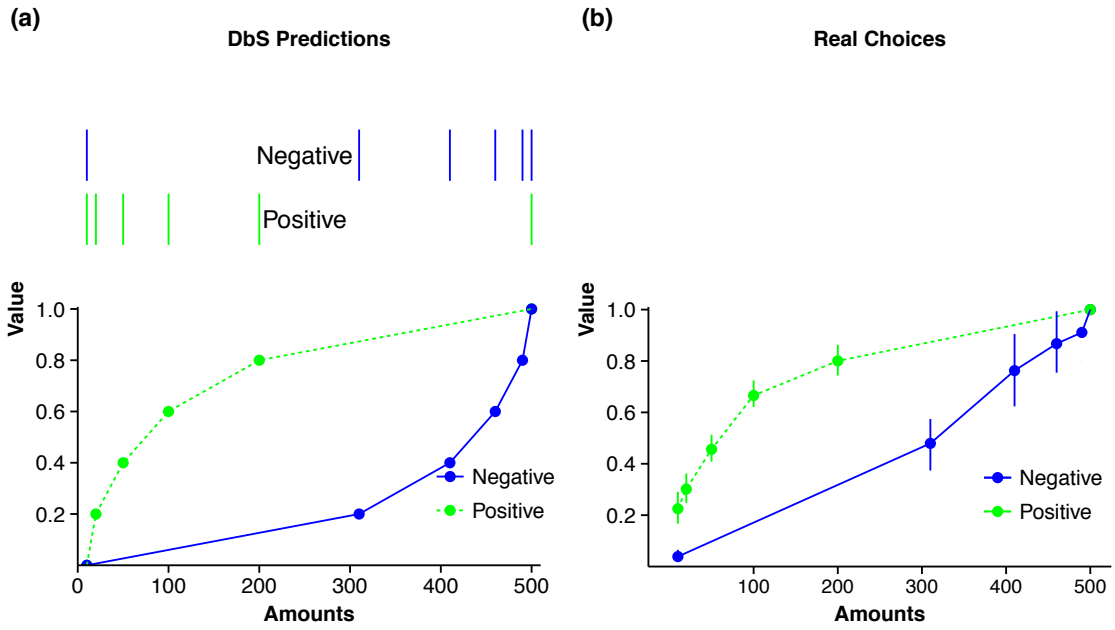

Figure 1.2. Figure 1.2(a) shows the predictions given participants experience the negative or the positive condition. The top part represents the distribution of amounts in the two conditions of the experiment. Figure 1.2(b) shows estimated value functions from real choices under the according conditions.

To test this hypothesis, Walasek and Stewart manipulated the range of possible gains and losses to spread from $20 loss to a $20 gain (or $40 loss to $40 gain), a $20 loss to a $40 gain or a $40 loss to a $20 gain. They then asked participants to accept or reject different offers composed of losses and gains, like a 50% probability to lose $10 and 50% to win $10. What does the range of losses and gains mean for the ranks of the amounts in the respective categories? If the gain-loss ranges are symmetrical, $10 has the same rank among losses than it has among gains. Because the evaluation is predicted to be rank-dependent, participants should equally likely accept or reject a gamble offering to equally likely win or lose $10. If losses reach up to $20 and gains up to $40, then $10 has a higher rank among losses than

among gains. This means that participants should reject this gamble, because a high-ranking loss cannot be compensated for by a lower-ranking gain. The reverse is true for the condition where losses range up to $40 and gains only up to $20. Now, $10 ranks higher among gains than among losses. The gamble should be attractive, because the gain has a higher probability of favorable comparisons with the other gains than the loss has of favorable comparisons with the other losses. The authors found the predicted pattern: They demonstrated that just by manipulating the range of available outcomes, participants chose as if they were loss-neutral in the symmetrically distributed cases, loss-averse when the range of gains was larger than the range of losses, and even the opposite of loss-aversion in the reverse case.

Both these sets of experiments show a serious departure from and present a big challenge to accounts of decision making and judgment that assume an inherently stable mapping between objective values into their subjective equivalents. The formulation of models that allow for experiments considering different components of the decision making process will help in identifying the main drivers of valuation. In this thesis, we test ideas that all assume comparison processes as key to judgment and choice. All experiments presented are inspired by the finding that context strongly influences what we choose and like. As many of these findings can be accounted for by DbS, we use this model to identify potential mechanisms leading to the above mentioned contextual effects. This thesis comprises of five chapters, in which we test different aspects of the decision by sampling model using experiments, cognitive modeling and meta-analyses of our findings.

Chapter 2 and 3 are based on the findings from Stewart, Reimers, and Harris (2015). The aim was to test how reliable the rank effects mirrored by the difference in psychoeconomic functions (the SRH effect) are and to what degree they are

attributable to DbS proposed mechanisms. We used different within-subjects designs to measure differences in psychoeconomic functions: In a first step, we flagged gamble pairs, so that they belong either to a positively or negatively skewed distribution, to assess whether people can keep track of different samples of attribute values (as they are assumed to keep track of gains and losses separately). We then retrospectively divided the choices according to their flags and estimated the differences between the psychoeconomic function for the two different distributions of attributes. If participants do separate choice pairs according to their flags into different samples (just as assumed in Walasek and Stewart (2015), where gains are compared to other gains and losses to other losses), differences in psychoeconomic functions should resemble those estimated with a between-subjects design. In a second step, we unflagged those pairs and again retrospectively estimated those differences between psychoeconomic functions to check if the observed differences are indeed attributable to DbS-type processes. In Chapter 3, we investigated the role of working memory by using cognitive load manipulations, where participants made some choices under high cognitive load or no/low cognitive load, and again estimated the differences between psychoeconomic functions. If participants' preferences are formed via pairwise comparisons of attributes from memory and DbS is correct in asserting a role for working memory in making a series of binary ordinal comparisons, we should see a change in people's choices when working memory is loaded. In Chapter 4, we first reviewed the literature showing rank effects in risky choice and then present yet unpublished data by Stewart and Reimers (2008), where we estimated and compared effect sizes of attribute distributions on choice reversals and attractiveness ratings of gambles, which are also predicted by DbS. The experiments in Chapter 5 were set up analogously to the experiments in Chapter 4.

Here we moved from choice to valuation and were interested whether positive affect after winning a prize in light of a salient alternative and other foregone prizes would have an impact on positive affect ratings, yielding a comparable effect as in Chapter 4. Finally, in Chapter 6 we examined potential causes for the mutable-zero effect, which describes the phenomenon that an option is favored when it is described with a "pay zero" attribute instead of a "receive zero" attribute. We investigated how different salient attribute values and counterfactuals triggered by the words "pay" and "receive" could lead to such context effects, prompting choice reversals. Chapter 7 concludes, and describes implications for future investigation.

2 Examining how utility and weighting functions get their shapes: A multi-level, quasi-adversarial, replication

Recently, Stewart, Reimers and Harris (2015, SRH hereafter) presented evidence from a series of experiments putatively demonstrating that the utility and probability weighting functions revealed by fitting standard economic models to binary choice data were sensitive to changes in the distributions of payoffs and probabilities in the choice sets. While for some the existence of such sensitivity may be no surprise (e.g. Drichoutis and Nayga, 2013; Etchart-Vincent, 2004; Fehr-Duda et al. 2010, 2011), the extent of malleability identified by SRH is considerable. For example, for some distributions of probabilities and payoffs, SRH were able to produce concave utility functions and inverse-S shaped probability weighting functions as commonly reported elsewhere in the literature; yet, for other distributions they were able to generate the mirror image patterns (i.e., convex utility and S-shaped probability weighting functions). As a convenient label, we will refer to the apparent malleability of the utility and probability weighting functions identified by SRH as the *SRH effect*.

At face value, the SRH effect poses a severe challenge to *any* model of risky decision making in the preference-theoretic tradition which, thereby, seeks explanations of choice grounded on the presumption of stable preferences. If a researcher can, as SRH explicitly suggest, choose the shapes of the functions they wish to reveal by adjusting the set of gambles used to elicit them, then the interpretation that such procedures reveal underlying preferences is undermined and central premises of welfare economics should be reconsidered. Hence, the SRH effect provides powerful new ammunition for those critical of the adequacy of preference-based models of risky-choice (Friedman et al. 2014, Gigerenzer, 2016). By contrast, the SRH effect provides support to those who favor process based models, and in

particular, it provides support for the model of Decision by Sampling (Stewart, 2009; Stewart et al. 2006), because predictions of this model (which we refer to as DbS for short) prompted discovery of the SRH effect.

But before interpreting the SRH effect as a strong challenge to preference based models (or support for procedural models including DbS), it is appropriate to question whether the effect is replicable and robust. That question is pertinent, not least, in the light of contemporary controversy surrounding the replicability of many of the findings in the behavioral sciences and elsewhere (e.g., Camerer et al. 2016; Maniadis et al. 2014; Open Science Collaboration, 2015). Given this background controversy and the challenging nature of the SRH findings, we believe good scientific practice demands careful scrutiny of the SRH effect, via attempts at replication, to properly assess its significance. With this motivation in mind, this paper reports an extensive set of replication experiments investigating two primary issues: first, we examine whether the SRH effect is replicable and robust to variations in experimental design; second, since we do find support for the SRH effect, we also probe its origins.

In what follows, we report a set of 14 new experiments conducted as part of what we call a *quasi-adversarial collaboration*, and combine these with a reanalysis of the fixed original experiments. The term "adversarial collaboration" has been used to refer to experimental research projects jointly planned and executed by two or more researchers (or research groups) who have ex-ante conflicting hypotheses about its outcome (for discussion and examples see Bateman et al. 2005; Corrigan, 2011; Kahneman, 2003; Latham et al. 1988; Mellers et al. 2001). While our collaboration does not have exactly this form (hence the qualifier 'quasi'), the seven researchers involved in this collaboration come from different disciplines (economics and

psychology), different labs, and have very different degrees of prior investment in the competing theoretical frameworks that would be supported or challenged by the existence of the SRH effect. We also use the qualifier "quasi' to signal that the set of experiments reported here did not emerge from a common plan of adversarial collaboration agreed before any of the experiments began. Instead, our collaboration emerged as subsets of the present co-authors began to discover that we were undertaking very closely related work exploring the SRH effect, independently, at different labs. We then began to compare results and, later, discuss designs for new experiments. Further experiments were subsequently run by different sub-groups of us, based in three different labs at two universities, using a mixture of lab-based and online protocols. The development of the designs involved varying degrees of consultation between us, as well as key variations in designs and procedures, which we document below. Through this process we have generated a rich source of evidence relative to the SRH effect, which we bring together in this paper.

The somewhat organic evolution of the collaboration does not mean that the set of replications, when viewed as a whole, lack structure. We will argue that, although we did not set out with this explicit purpose, the resultant set of experiments reflect and, indeed, extend a replication strategy proposed by Levitt and List (2009). They advocate a methodology involving replication at three levels: reanalysing data from the original study to be replicated; running fresh experiments using designs approximating the experiment to be replicated and thirdly, conditional on replicating the original results, running experiments to probe origins of the phenomenon observed. Our experiments involve replication at all three levels but we also add a further dimension to our analysis. Through our experiments, we generated a rich data set based on decisions of 1880 subjects which we use to run a meta-analysis. This

complements the individual experiments by placing confidence bounds on the size of the SRH effect and allowing assessment of how it varies with some key design features of our replications. As such we interpret part of our contribution as piloting an extended, four level, version of the Levitt and List (2009) methodology enhanced via a meta-analysis.

We reach two main conclusions. First, based on replication Levels 1 and 2, we conclude that the SRH effect is replicable, robust and the meta-analysis (Level 4) confirms a non-zero effect size. Hence, we reconfirm the challenge posed by SRH to those who seek to understand choice through the lens of standard preference based models. Second, based on the Level 3 analysis, we cast doubt on DbS as an account of the SRH effect since this model at best explains only about half of the effect size observed in our analysis. Hence, we identify the need for further investigation to explain the causes of the SRH effect that we observe in our data.

The paper is organized as follows. We first review key features of the original SRH study and its basic findings. We then summarize the analysis of our results from all four levels of the replication process. Next we discuss our main findings, and the last section concludes.

SRH's (2015) study: setup, motivation, methodology and results

The main results in the SRH original study are based on data generated from experiments in which individuals had to make a series of choices between pairs of gambles of the form "p chance of x, otherwise nothing" or "q chance of y, otherwise nothing" (where $p<q$ and $y<x$). SRH use these data to estimate utility functions over

monetary payoffs (x, y) by fitting an expected utility model and probability weighting functions over probabilities (p, q) by fitting a subjective expected utility model.[1]

All experiments followed the same logic whereby, for any given treatment, a fixed set of (five or six) money amounts was fully crossed with a fixed set of (five or six) probabilities to create a set of gambles. This was then used to generate a set of pairwise choices, for any given treatment of the experiment, comprising all possible non-identical and stochastically non-dominant pairwise choices from the full set of gambles. A further 30 choices were added in which one option stochastically dominated as a catch for participants paying insufficient attention to the task; participants who violated more than 10% of catch trials were excluded from the main analysis. The order in which choices appeared was randomized across participants.

There were two treatments in each experiment which varied according to either the skew in the distribution of money amounts, or the skew in the distribution of probabilities, used in construction of the choice sets. Each experiment then involved comparison of a treatment with a positive-skew distribution (of money amounts or probabilities) against a treatment with either a negative-skew or a zero-skew distribution (of money amounts or probabilities). It was comparisons between these pairs of treatments which generated the SRH effect.

The actual amounts and probabilities used in five of the different experiments reported by SRH are depicted in Table 2.1.

[1] There was an additional series of temporal discounting experiments, which we do not address in this paper.

Table 2.1

Outline of the amounts and probabilities used in each SRH original experiment to create the choice gamble.

Experiment	Manipulation	Skew	Amounts (£)	Probabilities (%)
Utility function manipulation				
SRH 1A	Amounts	Positive vs. Negative	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
SRH 1B	Amounts	Positive vs. Zero	10, 20, 50,100, 200, 500 vs. 0, 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
SRH 1C	Amounts	Positive vs. Zero	10, 20, 50,100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100
Probability weighting function manipulation				
SRH 2A	Probabilities	Positive vs. Negative	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	10, 20, 30, 40, 70, 90 vs. 10, 30, 60, 70, 80, 90
SRH 2B	Probabilities	Positive vs. Negative	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99

Notice that for the first three experiments in Table 2.1 (Experiments SRH 1A-1C) it is the distributions of the money amounts that vary between treatments, holding constant the set of probabilities used to construct the set of gambles. For example, in Experiment SRH 1A, the gambles are constructed using a common set of probabilities in both treatments (ranging from 20% to 100% in 20% steps); whereas the money amounts range from £10 to £500 in both treatments but for one treatment (with the positive-skew distribution) the intermediate outcomes are all in the lower half of the range, while for the other treatment (with the negative-skew distribution) all of the intermediate outcomes are in the upper half of the range. These experiments examined how changing the distribution of money amounts changes the revealed utility functions. In the last two experiments depicted in Table 2.1 (SRH 2A and 2B) the distribution of amounts is common across the two treatments for each experiment, but the distribution of probabilities changes between treatments. These experiments tested the same effect in the probability domain by changing the distribution of probabilities between treatments and examining the resulting probability weighting functions.

An example of the choice interface, based on SRH 1A, is shown in Figure 2.1. At the top of the screen participants saw the distributions of chances to win and prizes on offer across the set of choices. In most experiments, this information was on screen continuously while subjects made their decisions.[2] Although subjects were not explicitly informed about the number of choices they had to make, a bar at the bottom of the screen kept track of their progress. The number of the unique non-dominated pairwise choices was 150 for experiments SRH 1A, SRH 2B, 120 for experiments SRH 1B, SRH 2A and 100 for experiment SRH 1C. After making each decision, the

[2] This holds for all experiments but for SRH 1B, where subjects only saw the series of choices, without the distribution of chances and prizes on offer.

next choice appeared automatically. In any given choice, the two gambles were presented in the form of text in separate 'buttons' and subjects indicated their decision by clicking on one of them. They were told that at the end of the experiment one of their choices would be randomly selected and the gamble that they had selected in that choice would be played out for real using an exchange rate: 1 pound equals 1 pence.[3] All SRH original experiments were conducted at the University of Warwick (with SRH 1C run online).

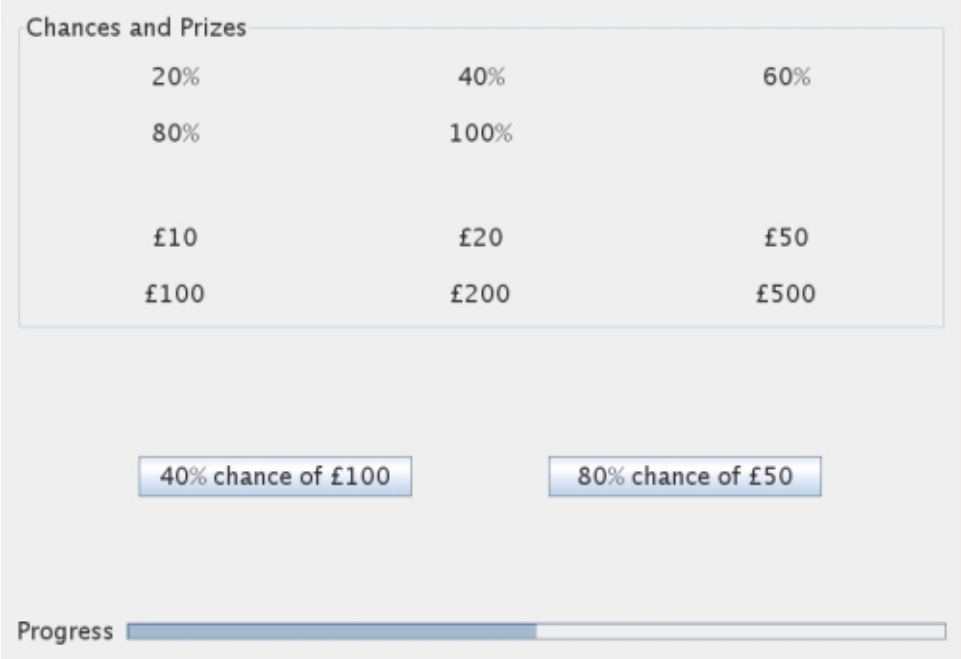

Figure 2.1. Interface used in SRH 1A.

In a moment we will review the main findings of SRH. As a prelude to that, we note that the treatment comparisons in SRH have a particular theoretical motivation because, as SRH explain, the DbS model predicts systematic differences between them. The DbS model is a mechanism for the construction of choices from a series of ordinal comparisons between pairs of attribute values. Readers interested in

[3] In Experiment SRH 1A two choices were randomly selected for payment. The exchange rate was halved for this experiment. In Experiment SRH 1C the choices were hypothetical rather than incentivized.

the details of the DbS model should consult Stewart (2009), Stewart et al. (2006), and SRH (2015). For now, the following property is sufficient: Because the probability that an attribute value will win an ordinal comparison is given by its rank position within those attribute values available, DbS predicts people will choose as if subjective value is the *rank position* of an attribute value within all those available. For example, consider the evaluation of the amount £200 in the context of the distributions used in experiment SRH 1B. In the positive-skew treatment where the amounts are £10, £20, £50, £100, £200, and £500, the £200 outcome is better than 4 out of 5 of the other outcomes that will be encountered. But now consider the evaluation of £200 in the context of the zero-skew treatment of experiment SRH 1B with the distribution of £0, £100, £200, £300, £400, £500. In this case, £200 is better than only 2 out of 5 other outcomes that will be encountered. Thus DbS implies a higher utility for £200 in the positive-skew treatment, but a lower utility for £200 in the zero-skew treatment.

Figure 2.2 shows the estimated utility and weighting functions for all five SRH original experiments. In line with DbS predictions, the utility functions in the utility experiments (in Figures 2a, 2b, 2c) and the weighting functions in the probability experiments (in Figures 2d, 2e) were more concave when the skew in the distribution of amounts and probabilities was positive than when it was negative or zero. For example, the utility for £200 in SRH 2B is lower in the zero-skew treatment than in the positive-skew treatment. The intuitive explanation from DbS is that, because subjective value is given by rank position, subjective value must increase most quickly where attribute value densities are highest. This means a steeper increase early on distributions with positive-skew compared to distributions with negative- or zero-skew.

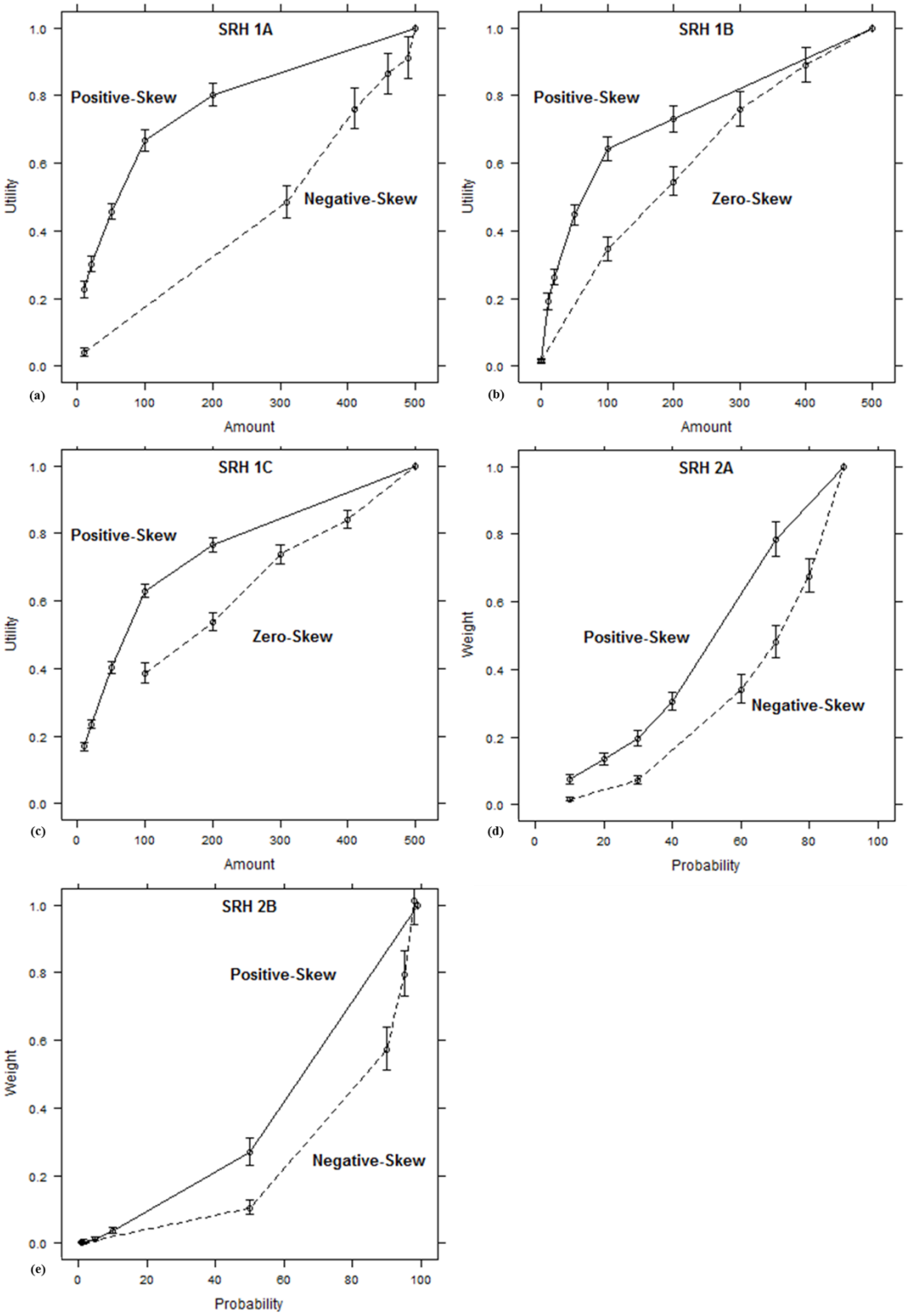

Figure 2.2. The revealed utility and probability weighting functions from SRH. Error bars are 95% confidence intervals. *Source*: Adapted from SRH (Figures 4a, 4b, 4c, 5a, and 5b respectively).

Four Level replication

We now present the results of our new analysis, which as explained above, involves four levels of replication, extending the methodology advocated by Levitt and List (2009) with an additional stage of meta-analysis to take advantage of the rich data set we have generated via our new experiments. In terms of data analysis, our strategy is to use methods which are essentially the same as those applied by SRH, except for some refinements explained below. We then hold constant the statistical methods that we apply across the four levels of replication reported in this study.[4] Analysis for each of the four levels is presented as a separate subsection.

Level 1: Replication by reanalyzing the original data

As a first step we reanalyzed SRH's original experimental data and estimated the revealed utility and weighting functions. Specifically, we fit an expected utility model with a Luce (1959)-Shepard (1957) choice rule incorporating a stochastic component to estimate utilities.

$$Prob(Choose\ safe) = \frac{bias\ (q\ u(y))^{\gamma}}{bias\ (q\ u(y))^{\gamma}+(p\ u(x))^{\gamma}} \quad (1)$$

Here $u(y)$ is the utility of amount *y* and $u(x)$ is the utility of amount *x, q* is the gamble probability of winning *y* and *p* is the gamble probability of winning *x*, $bias$ is a general tendency to choose safe irrespective of the actual amounts and probabilities on offer, and γ controls the level of determinism in responding (γ =1 gives choice probabilities proportional to the expected utilities, and γ >1 gives more extreme choice probabilities, so gambles with a slightly higher expected utility are very likely to be chosen).

[4] While one could entertain different approaches to modelling, basing our approach on that used by SRH minimizes the chance that differences between our results and theirs are due to modelling differences; and holding the approach constant within our analysis allows us to rule out the possibility that differences across our experiments or levels of replication could be plausibly attributed to our statistical modelling methodology.

The advantage of this model, though it is perhaps not obvious, is that for simple gambles it can be estimated as the following logistic regression:

$$\log\left[\frac{Prob(Choose\ safe)}{1-Prob(Choose\ safe)}\right] = \nu + \omega \log\left(\frac{q}{p}\right) + \sum_i \beta_i\ X_i\ . (2)$$

In Equation 2, each X_i is a dummy variable indicating the presence of amount *i* as *y* (coded +1), as *x* (coded -1), or absent from the choice (coded 0); $\nu = \log(bias)$; $\omega = \gamma$; and the utility of each amount $u(i) = \exp(\beta_i/\gamma)$. SRH give details in their Appendix A. A corresponding logistic regression for estimating weights for probabilities is given by exchanging the roles of *p* and *q* and of *x* and *y*.

$$\log\left[\frac{Prob(safe)}{1-Prob(safe)}\right] = \nu + \omega \log\left(\frac{y}{x}\right) + \sum_i \beta_i\ X_i \quad (3)$$

where the weighting of each probability is given by $w(p_i) = exp(\beta_i/\gamma)$.

Our analysis differs from SRH's original analysis in terms of the estimation of the confidence intervals. We used a more reliable bootstrapping method instead of the standard errors from the model fits as in SRH, because we wanted to allow for the possibility of asymmetric confidence intervals. Furthermore, we estimated the revealed functions separately for each condition. SRH mistakenly used one model for both conditions, but this is incorrect; random effects cannot be estimated for all amounts for all participants, because each participant experienced only a subset of amounts. Level 1 replication identified some trivial calculation errors in SRH's original analysis, but does not change the conclusions of the original paper. A detailed comparison between SRH's original analysis with and without those calculation errors can be found in Appendix A.

The revealed functions from the replication analysis are shown in Figure 2.3. The panels correspond to those in Figure 2.2, where the original experiments are presented. For example, a direct comparison for Experiment SRH 1A shows that the replicated functions depicted in Figure 2.3a strongly mirror the original functions

depicted in Figure 2.2a. The same result holds for all experiments. Hence, we conclude that Level 1 analysis successfully replicates the SRH effect, both in the utility experiments and in the probability weighting experiments.

Level 2: Replication by using a new subject pool

In this section, we report the results of a series of eight new experiments run by different subsets of the current authors. Each experiment was designed to replicate one of the original SRH experiments but with several design variations introduced across our studies as set out below. Table 2.2 presents the details of the Level 2 experiments. For example, at the top of the table, the experiment labelled "L2.a" is a replication of the original experiment SRH 1C comparing positive-skew and zero-skew distributions of money amounts. The distributions of money amounts and probabilities correspond exactly with those in SRH 1C. Subsequent columns of Table 2.2 indicate that L2.a was an incentivized experiment conducted with 54 incentivized participants at University 1 using 75 randomly selected choices from the original study. Looking further down Table 2.2, you will see that we have replicated examples of both the utility and the probability weighting experiments, though there is a focus on the utility experiments. This is partly an accident of history reflecting decisions made in the different labs when they started running these experiments independently. But, a focus on the utility dimension may, nevertheless, be useful for several reasons. First, utility is arguably a more fundamental concept, compared to probability weighting, in models derived from the preference theoretic framework; it features in a wider class of models and it is the core subjective dimension in what has been the leading theory of risk preference–that is, expected utility theory.

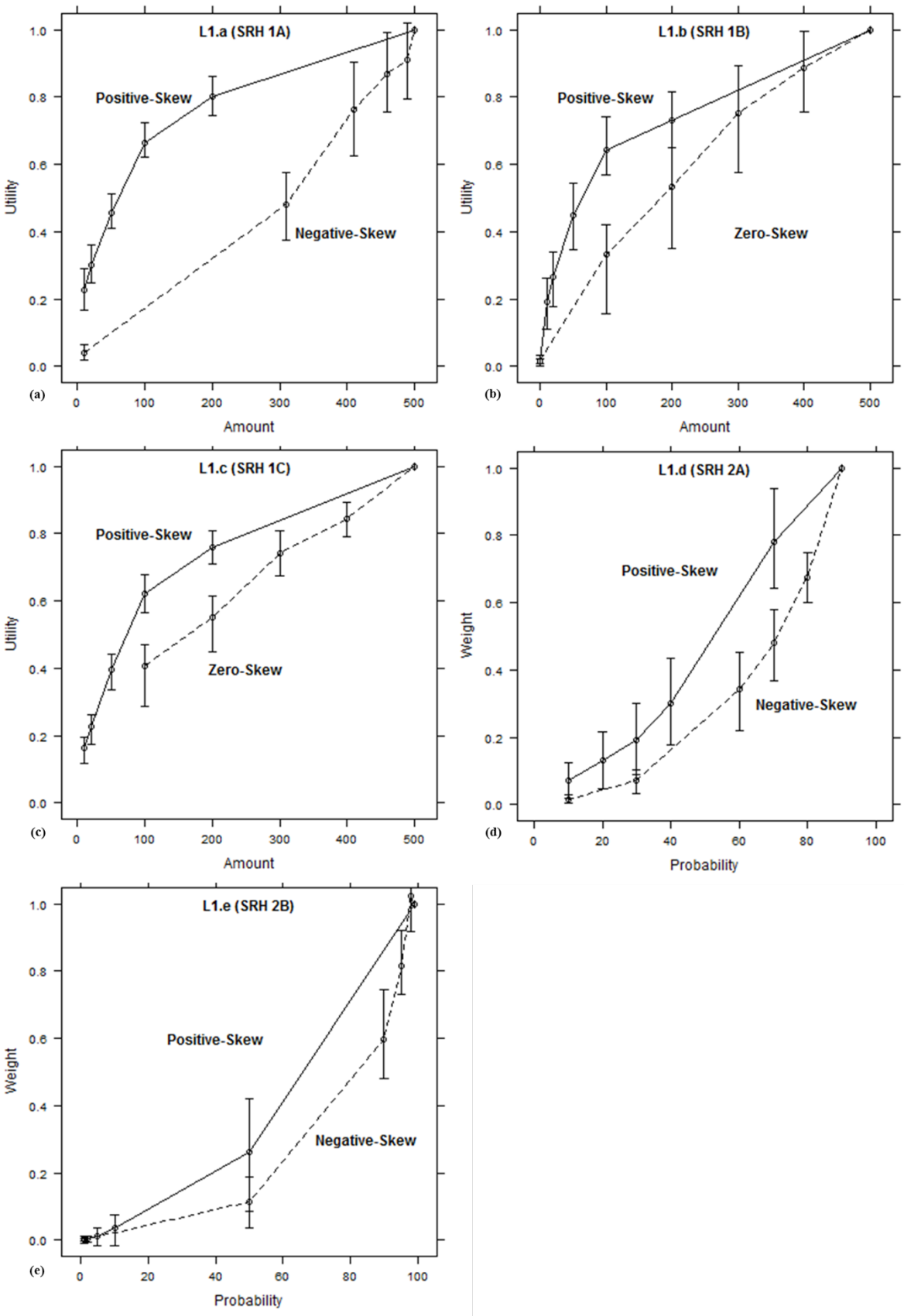

Figure 2.3. The revealed functions obtained from the replication analysis. Error bars are 95% confidence intervals.

While this focus seems justified by these arguments, we also wished to include at least one Level 2 test in the probability domain – hence the inclusion L2.h (additional experiments reporting manipulations of probability are reported as part of Level 3 replications). Notice that across the series of experiments there is variation in: the skew (positive- vs negative- or zero-skew), the domain (utility or probability), the location (group that conducted the experiment), whether the experiment was conducted online or in the lab, the number of participants, the number of trials and the incentives (by using both incentivized and hypothetical experiments). We think there is some advantage in focusing on particular SRH experiments to see the effects of small changes in procedures holding constant the distributions of amounts and probabilities. For this purpose we used SRH 1C, which contrasted positive-skew with zero-skew distributions of outcome values making experiments L2.a-L2.e different replications of SRH 1C. The advantages of choosing to replicate mainly the experiment SRH 1C are twofold: First, we can use the difference between the utilities of the common amounts for a direct comparison (there are no common amounts between the positive-negative conditions); Second, the round amounts used in the zero-skew condition are more representative of amounts often experienced by subjects in other experiments.

We also replicate SRH 1A, comparing positive-skew and negative-skew in the distributions of amounts in a different lab than SRH 1A (our experiment L2.f). In experiment L2.g we pushed the boundaries further by creating a distribution different from all the experiments reported in SRH and collected the data online. Finally, L2.h is a replication of SRH 2B and examines the probability domain.

Table 2.2

Outline of the properties of all replication experiments from Level 2.

Replication (Original)	Skew	Domain	Amounts (£ or $)*	Probabilities (%)	Location (Sample)	N[†]	No. Trials[¥]	Incentives
L2.a (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Student sample)	54	Mean of 75 randomly selected	Course credit plus up to £5
L2.b (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Prolific Academic)	200	150	£1.80 (non-incentivized)
L2.c (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	492	40	$1.80 (non-incentivized)
L2.d (SRH 1C)	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	145	Mean of 75 randomly selected	$1.80 (non-incentivized)
L2.e (SRH 1C)**	Positive vs. Zero	Utility	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (50% MTurk and 50% Prolific Academic)	183	150	$2.25 or £1.50 and up to $/£25

L2.f (SRH 1A)**	Positive vs. Negative	Utility	10, 20, 50, 100, 200, 500 vs.10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	40	180	£0 up to £5
L2.g (New distribution)	Positive vs. Negative	Utility	5, 10, 20, 50, 100, 200, 500 vs. 5, 300, 400, 450, 480, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (MTurk)	154	180	$3 (non-incentivized)
L2.h (SRH 2B)**	Positive vs. Negative	Probability	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99	University 2 (Student sample)	29	180	£0 up to £5

Notes: *We used $ for all Amazon Mechanical Turk (MTurk) samples and £ for all Prolific Academic and student samples. In experiments L2.f and L2.h, where the experiment was conducted using z-tree (Fischbacher, 2007) and subjects were recruited via the online system ORSEE (Greiner, 2015), a show-up fee of £7 was added to the earnings from the experiment.

[‡] Catch trials were not included in replication experiments L2.b, L2.c, L2.d, L2.e and L2.g. 2 subjects violated dominance in more than 10% of catch trials in experiment L2.f and 1 in experiment L2.h. These subjects were excluded from further analysis. In experiment L2.e we decided in advance to take the conservative approach of removing people in the 5% of fastest or slowest people, all multiple submissions from the same IP address, and the 5% of people who alternated the most or the least between left and right responses. We removed 56 participants based on the above criteria that are not included on the reported sample size.

[¥] In experiments L2.a, L2.c and L2.d we ran fewer trials to make the duration of the experiment shorter. This allowed us to run the experiment with as many participants as possible within a fixed budget

**These are experiments where I, Emina Canic, was not involved in up until the analysis of the data

To estimate the utility and probability weighting functions, we followed the same procedure as in Level 1; that is, we fitted Equations (2) and (3) to the choice data for the utility and the probability experiments respectively. Figure 2.4 shows the revealed utility and weighting functions from the set of Level 2 replication experiments. For example, the top left panel of Figure 2.4 depicts the revealed utility functions from experiment L2.a, where the resulting functions were more concave in the positive-skew treatment compared to the zero-skew treatment: Subjects assigned higher utilities to the common amounts of £100 and £200 when they experienced them in the positive-skew treatment (relative to the zero-skew treatment).

Eyeballing of Figure 2.4 reveals that the SRH effect is replicated in seven of the eight Level 2 experiments: in all cases apart from experiment L2.e, we see a more concave utility function in the positive-skew treatment compared to the other (negative-skew or zero-skew) treatment. While we cannot rule out other interpretations, it seems possible that the failure to replicate in L2.e may be due to random error. Either way, however, this overall pattern of results is strong evidence that the SRH effect is replicable and robust to a range of changes in procedures and subject pools.

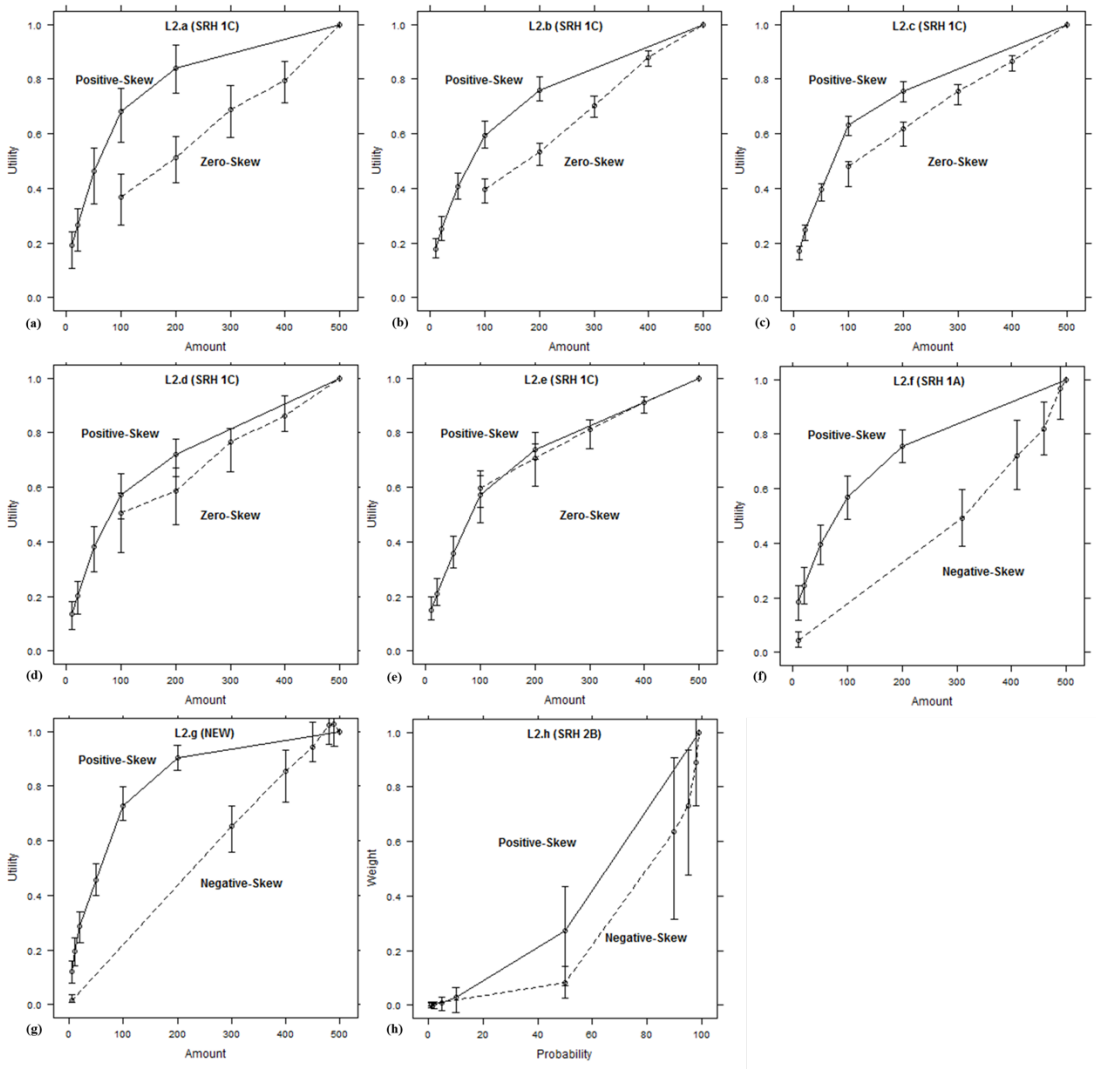

Figure 2.4. Revealed functions from the replication experiments in Level 2. Error bars are 95% confidence intervals.

Level 3: Replication by implementing a new design

While our Level 2 analysis provides substantial evidence suggesting robustness of the SRH effect, it does not show whether the conceptual interpretation of the original finding is correct. We therefore proceeded to a form of analysis in the spirit of Level 3 in Levitt and List's (2009) taxonomy. They suggest that Level 3 replication should involve creating alternative experimental designs, in order to test the same hypothesis as that tested in the original target of replication. Our strategy is to develop new designs intended to test the explanation of the SRH effect based on

DbS. We do this in two ways: first by running new experiments in which, conditional on the truth of the DbS account, it may be harder for the SRH effect to work; second by running tests in which, conditional on the truth of the DbS mechanism, we will eliminate the SRH effect altogether.

To this end, in one series of Level 3 experiments, which we refer to as 'flagged' experiments, we used a within-subjects design and presented participants with choice options originating from two different choice sets, one with a positive-skew distribution and the other with a negative-skew distribution. There are two key design changes in these flagged experiments (relative to Level 2 experiments): one is that *individuals experienced both choice sets within the same task*; the other is that *gambles from the different choice sets were flagged* to indicate which gamble pair belonged to which choice set. For example, in one of the experiments different sets of choices were described by different people using video recordings. One speaker "Joanne" was in her early twenties, white British woman; the other, "Patrick" was in his late twenties, white North American man. On a given trial, both of the gambles presented were drawn from the same set; that is, either both were from the set with positively skewed rewards or both were from the set with negatively skewed rewards.

For this within-subjects design, all option pairs from the two different distributions were merged to compose one choice set and this was presented to subjects in a randomized order. When it comes to analyzing the data, however, we retrospectively split the data, by choice set, to test whether there is an SRH effect observable (even though decisions for the two types of choice set were made by the same individuals). What would the model of DbS predict for such tests?

If participants simultaneously keep track of the two separate distributions (one presented by Patrick, the other presented by Joanne), and choose as if their decisions

are informed by applying DbS separately to choices that come from the separate distributions, then DbS would predict that splitting the combined choice set into the original two sets for the analysis will yield the same effects in the flagged within-subjects experiments as in the original between-subjects experiments. On the other hand, we do not know that individuals would try to use flags and separate the distributions accordingly. Even if they did it seems possible that SRH effects might be reduced due to interference and imperfect memory. As such our flagged experiments provide a tougher environment for the operation of the mechanisms imputed by DbS.

Our second series of Level 3 replication experiments (which we label "non-flagged") uses the same within-subjects design just described, except that in these experiments *we provide no flags*. Hence, in these experiments participants had no way of knowing what distribution the choices belonged to and had therefore no way of attributing choices to one distribution or another. As such, while a DbS model would still imply that any measured utility and probability weighting function would depend on the background distributions of probabilities and amounts, in this case there is effectively only a single background distribution provided in the experiment. Yet, as experimenters, we can still retrospectively split the data according to the two different distributions generating the choices for the analysis. According to DbS, the SRH effect should disappear. Should we continue to observe the SRH effect this would be evidence against DbS's proposed mechanism as the cause of this context effect. In fact, it would be a finding that no existing model could account for.

Because of the completely within-subjects nature of the experiments, we reverted to the original SRH modelling, where we fit one model to both conditions in an experiment. Random effects can now be estimated within one model, because all participants experience the attribute values from both distributions. In the logistic–

regression–form the model looks as follows for the experiments estimating a utility function:

$$\log\left[\frac{Prob(Choose\ safe)}{1-Prob(Choose\ safe)}\right] = \nu + \tau cond + \omega \log\left(\frac{q}{p}\right) + \xi cond \log\left(\frac{q}{p}\right) + \sum_i \beta_i X_i \quad (4)$$

An addition to the model described in Equation 2 are the terms involving *cond. Cond* is a dummy to account for the two conditions, which set to 0 indicates the positive-skew and set to 1, the other condition. Setting

$\log(bias_{cond}) = \nu + \tau cond$, $\gamma_{cond} = \omega + \xi cond$, and $u_{cond}(amount_i = \omega + exp(\beta_i/\gamma_{cond})$ follows, when Equation 1 is adapted to account for both conditions.

To estimate the probability weighting functions we fit the following model:

$$\log\left[\frac{Prob(safe)}{1-Prob(safe)}\right] = \nu + \tau cond + \omega \log\left(\frac{y}{x}\right) + \xi cond \log\left(\frac{y}{x}\right) + \sum_i \beta_i X_i \quad (5)$$

and calculate the weights via $w_{cond}(p_i) = exp(\beta_i/\gamma_{cond})$.

These are exactly the models estimated by SRH (see their Appendix A) and are extensions of our Equations 2 and 3. Table 2.3 summarizes the set of Level 3 experiments. As in Level 2, we varied the domain, the location, the incentives, the number of trials, the number of participants, and whether the experiment was conducted online or in the lab. For example, the second row of Table 2.3 describes the details of experiment L3.a, which used the positive- vs. negative-skew in the distribution of amounts from the original experiment SRH 1A, and flagged them by labelling prizes as either mobile phones or holidays. We present the main results of Level 3 analysis in Figure 2.5.

Table 2.3

Outline of the properties of all experiments from Level 3.

Replication (Original)	**Skew**	**Amounts (£ or $)**[±]	**Probabilities (%)**	**Location (Sample)**	**N**[≡]	**N Trials**[⊥]	**Payment**
Flagged Experiments[∅]							
L3.a (SRH 1A)**	Positive vs. Negative	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	48	330	Course credit (non-incentivized)
L3.b (SRH 1A)**	Positive vs. Negative	10, 20, 50,100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (Student sample)	42	160	Some for course credit some were paid £6 plus up to £5
Non-flagged Experiments							
L3.c (SRH 1A)**	Positive vs. Negative	10, 20, 50, 100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student sample)	45	450	Course credit (non-incentivized)
L3.d (SRH 1A)**	Positive vs. Negative	10, 20, 50, 100, 200, 500 vs. 10, 310, 410, 460, 490, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 2 (Student Sample)	50	450	£0 to £5
L3.e (SRH 1C)**	Positive vs. Zero	10, 20, 50, 100, 200, 500 vs. 100, 200, 300, 400, 500	20, 40, 60, 80, 100 vs. 20, 40, 60, 80, 100	University 1 (50% MTurk and 50% Prolific Academic)	89	140	$2.25 or £1.50 and up to $/£25

L3.f (SRH 2B)**	Positive vs. Negative	100, 200, 300, 400, 500 vs. 100, 200, 300, 400, 500	1, 2, 5, 10, 50, 99 vs. 1, 50, 90, 95, 98, 99	University 2 (Student sample)	49	390	£0 to £5

Notes: ±We used $ for all Amazon Mechanical Turk (MTurk) samples and £ for all Prolific Academic and student samples.In experiments L3.d and L3.f, where the experiment was conducted using z-tree (Fischbacher, 2007) subjects were recruited via the online system ORSEE (Greiner, 2015), a show-up fee of £7 was added to their earnings.

≡Catch trials were not included in replication experiments L3.b and L3.e. 1 subject violated dominance in more than 10% of catch trials in experiment L3.d and 3 in experiment L3.f. These subjects were excluded from further analysis. In experiment L3.e, we decided in advance to remove people in the 5% of fastest or slowest people, all multiple submissions from the same IP address, and the 5% of people who alternated the most or the least between left and right responses. We removed 32 participants on these criteria that are not included on the reported sample size.

⊥In experiments L3.b, L3.e we ran fewer trials with as many participants as possible within a fixed budget.

∅The flagged experiments were as follows: In experiment L3.a for half of the participants 150 positively skewed questions were framed as holidays and 150 negatively skewed questions were framed as mobile phones. For other participants it was the other way around; in experiment L3.b the choice options from the different distributions were described by different people using video recordings. For half of the participants, Patrick described 80 gambles from the positively skewed set and Joanne described 80 gambles from the negatively skewed set; for the other half, this assignment was reversed.

** These are experiments, where I, Emina Canic, was not involved in up until the analysis of the data

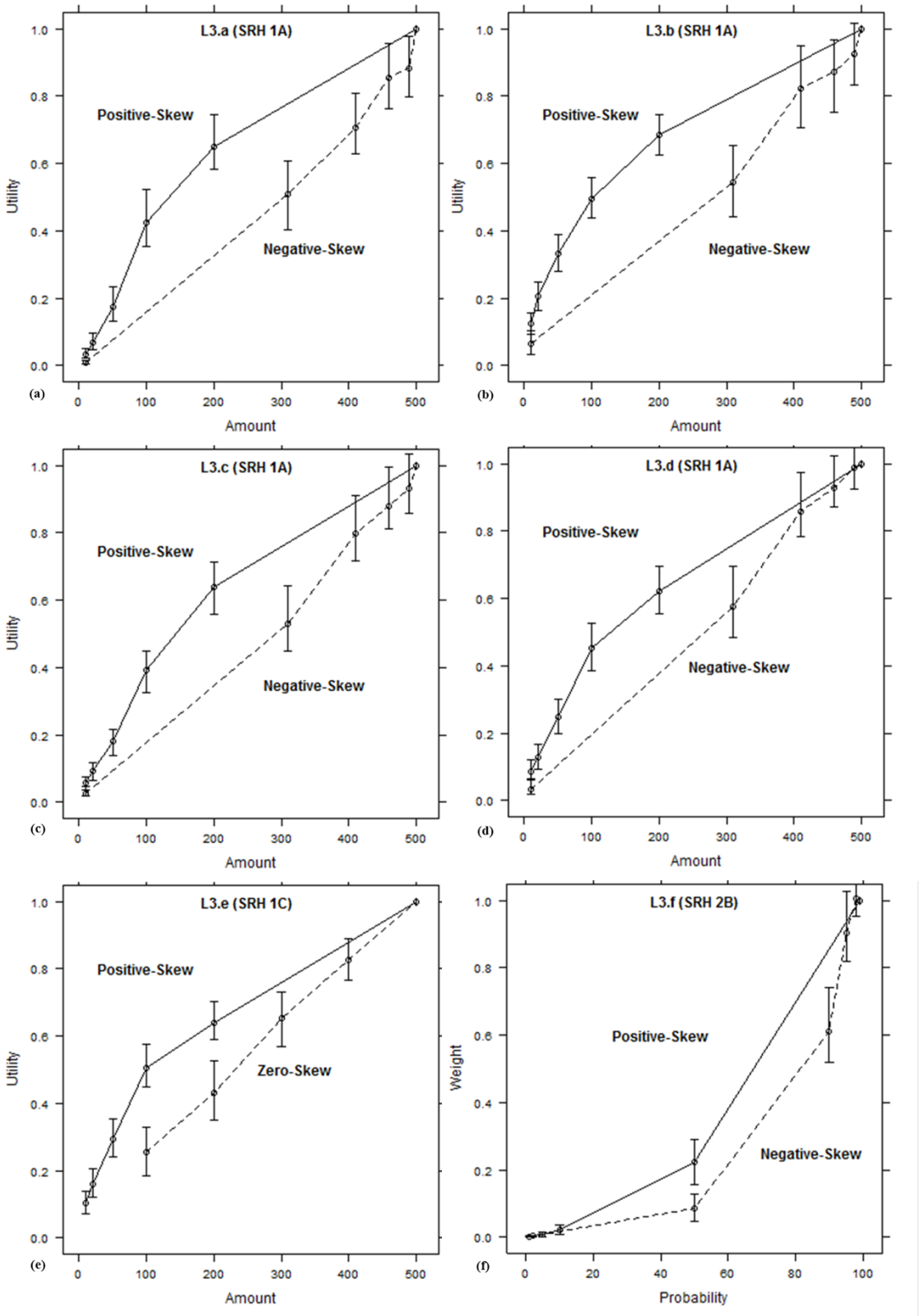

Figure 2.5. Revealed functions from the replications of SRH in Level 3 using a within-subjects design. L3.a-L3.b involve flagged choices, L3.c-L3.f do not. Error bars are 95% confidence intervals.

The top two panels of Figure 2.5 show results for the flagged experiments (L3.a and L3.b) where participants could potentially track what distribution the attributes in the choice set belonged to. The difference between the revealed utility functions for the two differently distributed samples of amounts remain: the utility function using the choices from the choice set with positive-skew is more concave than the one from the choice set with negative-skew.

Surprisingly–and in contrast to DbS predictions–when estimating the revealed utility and weighting functions in the non-flagged experiments L3.c-L3.f, the comparison of curves within each panel still shows the pattern of an SRH effect. It is hard to see how these differences can be rationalized with the DbS model. Hence, our conclusion is that Level 3 results at least partially challenge the DbS interpretation of the SRH effect.

Level 4: Meta-analysis

The nineteen experiments we have analyzed in this study (including the five Level 1 cases from the original SRH study) provide a large data set, based on the decisions of 1880 subjects, and are suitable for conducting a meta-analysis in order to examine the overall size, variability, and moderators of the effect of context on utility and probability weighting functions. We see this as a useful complement to the replication analysis of Levels 1 – 3 and so as a natural extension of the Levitt and List methodology, in cases where a suitable, comparably rich, data set has been generated.[5]

We first estimated the overall effect size using the differences in the revealed utility and weighting functions between conditions across all experiments for one attribute value. Subsequently, we calculated the effect sizes separately for (i) the between-subjects design experiments (ii) the within-subjects flagged experiments and (iii) the within-subjects non-flagged experiments. From an explanatory point of view, the comparison of results for experiments in categories (i) and (iii) is especially interesting: If the within-subjects

[5] We do not consider a meta-analysis of replication studies generated from a quasi-adversarial collaboration as a sufficient condition to obtain a definitive estimate of the underlying effect sizes. However, we regard it as an important step in the right direction.

experiments' non-flagged effect size were as big as the between-subjects experiments' effect size, we could reject the DbS interpretation of the context effects, because the effect in the non-flagged experiments cannot emerge through the mechanisms proposed by DbS. However, if the effect size in the non-flagged within-subjects design experiments is smaller, DbS could still remain a candidate model for the interpretation of some part of the SRH effect.

For the effect size measure, we identified the amount or the probability that was common to the two distributions in a given experiment and calculated the difference between the utility (or weighting) functions of the two distributions at that point. If no attribute was common within an experiment, we picked the two attribute values that were most similar to each other across the positive- and zero- or negative-skew distributions. For the experiments comparing zero- vs positive-skew, £200 (or $200) occurs in both distributions, so we subtracted the utility estimate of £200 in the zero-skew treatment from the utility of £200 in the positive-skew treatment. For experiments with positive- vs. negative-skew distributions of amounts, we calculated the difference between the estimated utility of £200 and the utility of £310 for the utility experiments (we used the £200 vs. £300 comparison for experiment L2.g). Note that DbS predicts £200 in the positive-skew condition to be a higher utility than £310 in the negative-skew condition. The fact that £310 is a higher number than £200 makes this a conservative comparison. For the experiments manipulating the probability distribution we calculated the difference between the estimated probability weights of the common 30% (for SRH 2A replications) and the common 50% (for SRH 2B replications) in each condition.

We fitted a linear random effects model to estimate the effect of experimental design (between vs. flagged within vs. non-flagged within) and the skewness comparison (positive-negative vs. positive-zero) using mean differences.[6] The results of the meta-analysis are depicted in Figure 2.6.

[6] The estimations were obtained by using the metafor package in R by Viechtbauer (2010).

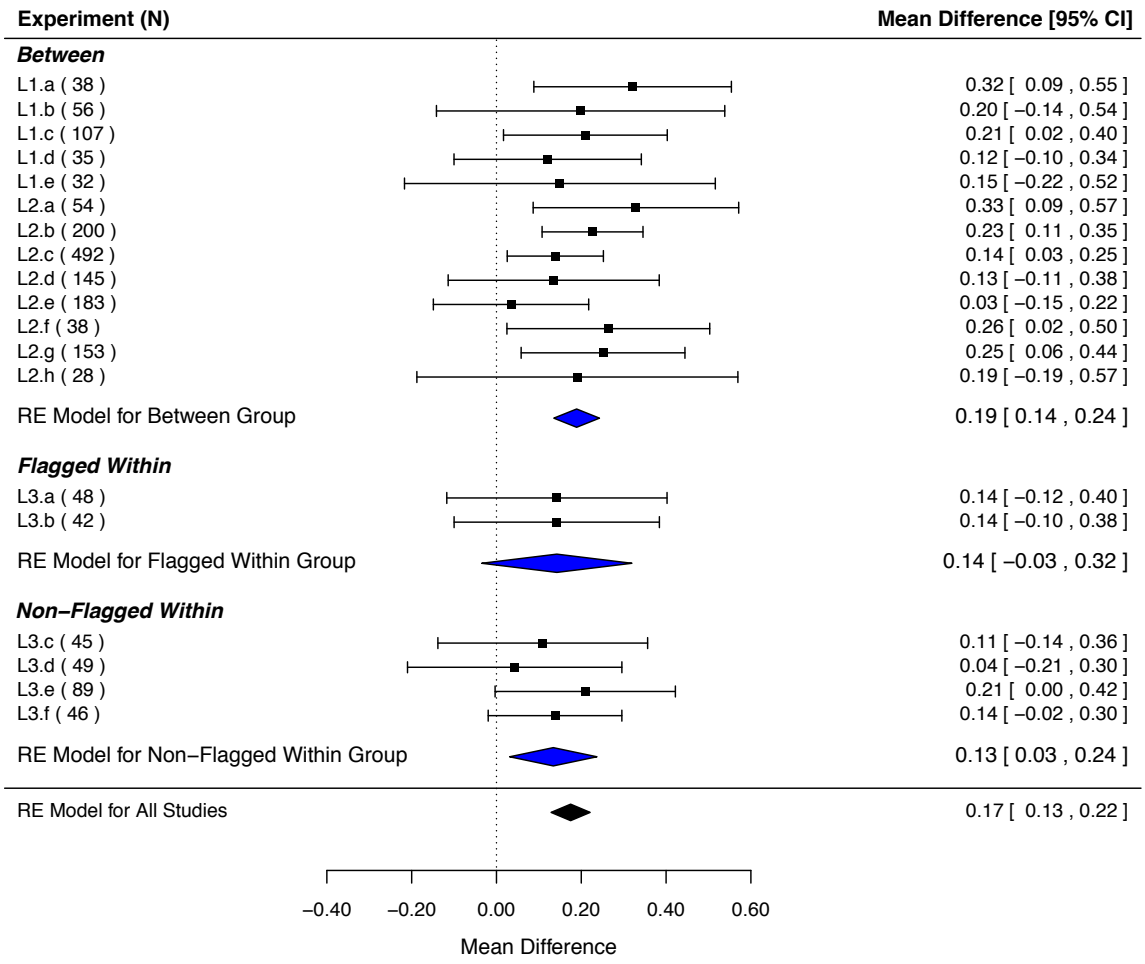

Figure 2.6. Meta-analysis results from Level 4. Mean differences (MD) and 95% confidence intervals are shown as a function of the experimental design for between-subjects, and for flagged and non-flagged within-subject experiments.

Figure 2.6 shows that our best estimate of the difference between the revealed utility or probability weight of the common attribute between the two distributions (positive-negative skew comparison or positive-zero skew comparison) overall is 0.17 95% CI [0.13, 0.22] on a scale where, arbitrarily, the utility of £500 for the utility experiments and the weight of 99% in the probability experiments are fixed at 1. Thus, taking all our experiments and SRH's experiments into consideration, the meta-analysis confirms that the SRH effect is replicable and real.

We estimated the effect of distribution comparison (positive-negative vs. positive-zero). The estimates do not differ between distribution comparisons, but could potentially

differ by approximately 0.10 in each direction, ($beta_{Distribution}$ = 0.00 95% CI [-0.10, 0.09]). That is, the difference between the estimated utility of £200 from the positive-skew condition and £200 from the zero-skew condition is similar to the difference between the utility of £200 in the positive-skew condition and the utility of £310 in the negative-skew condition. However, there might be slight differences as the confidence interval is reasonably wide, leaving open the possibility that there are real differences between positive- vs zero-skew comparisons and positive- vs negative-skew comparisons.

Looking at the effect sizes for the three designs separately, the effect is largest in the between-subjects experiments ($MD_{between}$=0.19 95% CI [0.14, 0.24]). The effect is reduced in the flagged within-subjects experiments ($MD_{flagged\_within}$=0.14 95% CI [-0.03, 0.32]) and is smallest–with a reduction of 30% of the between-subjects experiments–in the non-flagged within-subjects experiments ($MD_{non\text{-}flagged\_within}$=0.13 95% CI [0.03, 0.24]). According to the meta-analytic model, the differences could be very small, or they could be opposite, such that the effects are larger in the non-flagged experiments, or it could be that the difference between the between-subjects experiments and the non-flagged experiments is about as large as the effect itself in the non-flagged experiments, ($beta_{Design}$ = 0.04 95% CI [-0.04, 0.12]). Given that the DbS account cannot apply to the non-flagged experiments, and the effect size in the non-flagged experiments is estimated to be similar to the two other groups, it is likely that much of the effect in the flagged and between-subject experiments should not be attributed to the DbS model.

Conclusion

We showed that the qualitative effects of attribute distributions on utility and probability weighting functions reported by Stewart, Reimers and Harris (2015) are highly replicable: we obtained utility and probability weighting functions that were qualitatively very similar to those of the original study across a set of experiments (our Level 2 replications) involving various design changes. More substantial variations in the experimental design (our

Level 3 replication) plus our meta-analysis, however, suggest that although the SRH effect can be reliably reproduced, the interpretation proposed by SRH, based on the model of Decision by Sampling, is unlikely to be the *sole* explanation of the apparent manipulability of utility and probability weighting functions. The findings have two implications: First, they reinforce the challenge to preference-based accounts of risky choice behavior posed by the existence of the SRH effect; second, they suggest the need to pursue new directions in the hunt for an explanation of the SRH effect. One possible explanation for the SRH effect could be that the parameter estimates are biased in a way that the bias depends on the set of values that are presented to the participants. This would mean that the estimates for the uniformly distributed amounts and probabilities are biased in a different way from the estimates for amounts and probabilities that are either positively or negatively skewed. The consequence would be that the estimates between the conditions (with differently distributed attribute values) are not comparable and thus should not be compared, because the parameters are not generalizable across differently distributed attribute values. Stewart, Canic and Mullett (2017) provide a more detailed explanation of this account.

3 Choices remain context sensitive, even under high cognitive load

To what degree and in what way does cognitive load alter our decisions under risk? The main interest in this chapter lies in the effect of cognitive load on our risky decisions. When our working memories are filled up by a concurrent task (a cognitive load), working memory capacity is less available to support our risky decision making. The studies presented here will investigate whether our preferences, as measured by estimating the utility functions that best describe our choices, differ when we are choosing under high or low cognitive load. We contrast three accounts of how cognitive load may affect our risky decision making: cognitive load could lead to more or less risk averse choices, to more or less noisy choices, or to choices more or less sensitive to the context in which they are made.

We will start by reviewing the literature which supports the argument that reduced cognitive capacity will make us more risk averse. We first address the literature that shows how lower intelligence is associated with reduced cognitive capacity, and then the literature associating lower intelligence with more risk aversion. We then review a more recent reevaluation of this evidence that instead illustrates how reduced intelligence is associated with more noisy decision making. Finally, we introduce our new hypothesis, which emerges from the DbS model and the consideration of the role of memory in the decision making process. Based on the decision-by-sampling account cognitive capacity should reduce sensitivity to the decision context.

How do intelligence and working memory capacity relate to each other?

Making decisions under cognitive load implies that participants' working memory capacity at this point will be reduced. Research that ties working memory capacity (WMC) to general cognitive abilities, which in turn moderate decisions, including those in risky environment, makes the link between WMC and intelligence important. We will therefore start by reviewing research that investigates the relationship between WMC and general cognitive abilities.

In the early days, some researchers defined the connection between WMC and cognitive abilities to lie in the ability to keep a representation active, especially whilst distracted (Engle, Tuholski, Laughlin & Conway, 1999). Although the relationship between working memory capacity (WMC) and general cognitive abilities has been widely established, there is still an ongoing discussion about what the direct link between the two is (Unsworth & Engle, 2005). In contrast to more recent research, initially some thought that WMC was a construct that underlies intelligence or that directly influences it (Conway, Kane, & Engle 2003). Some researchers have even argued that WMC and general intelligence are equal (Colom, Flores-Mendoza, & Rebollo, 2003; Engle, 2002; Jensen, 1998; Kyllonen, 2002). More recently however, some of those have revised their argument (Conway, Kane, & Engle, 2003). To establish whether the two constructs were isomorphic, Ackerman, Beier and Boyle (2005) have conducted a meta-analysis and indeed have shown that general cognitive abilities are highly correlated with working memory capacity, that they share around a quarter of their variance, but that they are not the same thing. Whilst some researchers argued that a quarter is a huge underestimation and that the true share of variances is around 50 percent, they all agreed that the two constructs are not identical (Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhem, & Süss, 2005). Furthermore, Oberauer et al. clarified that the goal of investigating the correlation between WMC and intelligence is to "[…] validate WMC as an explanatory construct for intellectual abilities" (p. 64). This is important, because WMC is a central element in theories of cognitive architecture, and is proposed to play a crucial role when performing complex tasks, like making decisions. There is however, some controversy in regards to how similar or different the two constructs, intelligence and WMC, are. A big difference between the two is that whilst there are testable theories of what WM is and how it functions, no comparable theories exist for fluid intelligence. So far, we know that WMC (among other theoretical constructs) correlates most highly with general cognitive abilities and is a strong predictor of such (Martínez et al., 2011; Unsworth, Redick, Heitz, Broadway,

& Engle, 2009). Shahabi, Abad and Colom (2014) have recently investigated the relationship between short-term memory and fluid intelligence across different age groups and have found short-term memory to be a stable predictor (see also Barrouillet, Portrat & Camos, 2011). Because we have a clearer picture and more specific theories about WMC than we have about general cognitive abilities, learning more about WMC might help us understand what mechanisms underlie intelligent information processing.

The behavioral economists' perspective: In dual process theories cognitive capacity is associated with normative choices

The concept of how general cognitive capacities could be related to decision making is nicely captured in dual process theories of cognition. They make clear predictions about the effect of cognitive load on risky decision making. Dual process theories generally propose the existence of two systems that operate under different circumstances: an emotional system (System 1) that is responsible for fast intuitive decisions, which can be vetoed by the analytic, rational reasoning system (System 2), which serves to decontextualized and depersonalize problems. System 2 however, is cognitively costly and can therefore only be applied or can operate when the necessary cognitive resources are available. Cognitive load is therefore hypothesized to reduce the contribution from System 2, leaving System 1 unchecked or unmonitored. Dual process theories predict that the two systems lead to different responses (Stanovich & West, 2000). In terms of risky decisions, given a certain context one of the systems would predict risk seeking, the other risk aversion. Or, in terms of intertemporal decisions, patient versus impatient choices, respectively. This framework is also the one under which behavioral economists have made sense of many findings that demonstrate departures from subjective utility theory (Benjamin, Brown & Shapiro, 2013; Fudenberg & Levine, 2006; Mukherjee, 2010). The idea is that System 2 is an inner, rational economic man who, given sufficient time and resource, will overcome an irrational System 1, as popularized by Kahneman (2011).

In contrast to the traditional view in economics, behavioral economists have started taking into account cognitive abilities as individual factors that might moderate economic behavior. Barberis and Thaler (2003) argued that even on aggregate, behavioral biases found in their data cannot be eliminated by aggregating over everybody nor can they be eliminated by rational arbitrage (i.e., a process where "[…] rational [agents] can undo the dislocations caused by less rational agents […]" (p. 1052). Although classic economic models do not specify how cognitive abilities influence preferences, the last 10 to 15 years produced several empirical findings relating better outcomes in the labor market to higher cognitive abilities, and conversely relating excessive risk aversion and impatience to lower cognitive abilities.

Benjamin, Brown and Shapiro (2013) found that Chilean students with higher cognitive capacities are less likely to choose the safe option when low stakes are on offer and less likely to choose the sooner option when the time interval between the smaller sooner and the later larger reward is very short. Those children behave according to normative models. The authors found the same pattern when they manipulated cognitive capacity by asking some of their participants to memorize a long number, whilst others were not subjected to a memorization task. Memorizing a long number before making a choice led to more risk-averse decisions as compared to no memorization task before the choice. In a meta-analysis Shamosh and Gray (2008) found that higher cognitive abilities were positively correlated with patience. Bergman, Ellingsen, Johannesson and Svensson (2009) among others have found stronger anchoring effects in participants who scored lower on a cognitive abilities test. This difference was smaller for scores on the cognitive reflection test (see also Oechssler et al. 2009). On the other hand, using self-reported SAT scores as proxy for cognitive abilities, Stanovich and West (2008) could not establish an association between SAT scores and anchoring effects. Oechssler, Roider & Schmitz (2009) showed that behavioral biases like the conjunction fallacy and less accurate belief updating in probability estimation are more pronounced in people who scored lower on the cognitive reflection test, and that people's

probability judgments are impacted by working memory capacity (see also Frederik, 2005; Sprenger & Dougherty, 2006). Although higher cognitive abilities seemed to reduce these biases, they did not disappear. In addition, the observed differences in risk aversion and delay discounting between low scorers and high scorers were much smaller in Oechssler et al.'s study than in Frederik (2005) or Dohmen et al. (2007), although Oechssler et al. used real-staked lotteries and adopted a similar set of gambles as Frederik (2005) and Dohmen et al. (2007). Nevertheless, many other researchers too found that lower scores on a cognitive abilities test were related to more sooner lower payments and more risk averse choices (Bergman et al., 2010; Burkes et al., 2009; Payne, Samper, Bettman, & Luce, 2008).

Using dual system theories, economists argue that using System 2 would lead to rational, unbiased choices. People who do not use this analytical System 2 enough are just not smart enough. The rationale is that if only people were smart enough, they would use System 2 and they would make normative decisions like the ones predicted by classic economic models. Thus, descriptive as-if models like Cumulative Prospect Theory (CPT; Tversky & Kahneman, 1992), which say little about the actual cognitive mechanisms that underlie a decision, would be sufficient to predict everyday decisions normal people make under risk: The curvature of the value function and the S-shaped probability weighting functions would be good enough tools to capture the degree of irrationality. From this perspective, excessive risk aversion, loss aversion, or distortions of probability weighting are anomalies that only "idiots" show. (Although it is rational to be risk averse when high stakes are at play, Rabin, 2000, argues that the typical levels of risk aversion in low-stakes gambles in laboratory experiments are irrational because they imply absurd levels of risk aversion when scaled up to real-world problems.)

As long as people that differ in motivation or cognitive ability still exist, we can capture differences in their degree of risk aversion by estimating differences in the curvature of the utility functions: Assuming expected utility theory (EUT), people with lower cognitive

abilities should also choose as if they have a more concave utility function when making their decisions. This consequently leads to more risk averse choices. Whilst making everybody smart enough and giving them more time, would lead to everybody making more rational decisions–because they could perform expected utility calculations (Stanovich & West, 2000)–restricting cognitive capacities by imposing high cognitive load would lead to limits on the computations they could complete, and thus to less rational decisions.

In most cases so far, the impact of cognitive capacity on choice was studied by assessing cognitive abilities (i.e., measuring operation span or using the cognitive reflection test) and examining its moderating or mediating potential. However, to identify a potential causal relationship, one must manipulate cognitive load. On this basis, several studies have been conducted investigating to what degree anchoring effects, framing effects, self-control and delay discounting, dietary choice or risky choice is affected by high as opposed to no or low cognitive load. Whitney, Rinehart, and Hinson (2008) found that although the typical framing effects remained (participants are risk averse with a positive framing and risk seeking with a negative framing), under high cognitive load participants were more risk averse in both types of framing. Other researchers have shown that participants were more likely to eat unhealthy food over fruit and thus exhibit greater impulsivity when having to remember a seven-digit instead of a two-digit number (O'Donoghue and Rabin 2001; Shiv & Fedorikhin, 1999; Ward & Mann, 2000). Similarly, studies showed that under high working memory load people were less patient, preferring more sooner, smaller rewards and performed worse on a gambling task (Hinson, Jameson, & Whitney, 2002; Jameson, Hinson, & Whitney, 2004). In addition, Fudenberg and Levine (2008) have reported an Allais paradox reversal when they induced cognitive load. In a recent study, Deck and Jahedi (2015) used a within-subjects design looking at a number of interesting outcome variables. Their studies yielded effects of high cognitive load in terms of a reduction in numeracy, increase in risk averse choices,

impatience over money, but not over an unhealthy diet nor more anchoring within one participant.

Cognitive load is associated with more random responding, not more risk aversion

There is a very different interpretation of the findings reviewed above. Some studies point to an issue that was thus far completely overlooked when investigating the relationship between cognitive abilities and risky or delayed choices, despite the fact that there is older literature relating limits on computation to performance error across reasoning tasks (Cohen, 1991; Stein, 1996). The researchers argue that what everybody is observing and interpreting as an increase in risk averse or impatient choices, is actually an increase in randomness, due to mistakes when cognitive resources are insufficient. Franco-Watkins, Rickard & Pashler, (2015) as well as Andersson et al. (2013) provide compelling evidence that participants made more random choices under high cognitive load: Taking noise into account, the authors found no relationship between cognitive abilities and risk aversion. Andersson et al. argue that the reason why previous research found an association between risk and cognitive abilities was because those authors "[…] fail*ed* to account for the heterogeneous nature of the propensity to make mistakes, which may lead to biased inference about preferences for risk from observed choices (pp. 2-3)." What these authors found is that, if they ignored noise, participants with higher cognitive abilities were less risk averse. However, if noise was accounted for, the relationship disappeared and instead they found that participants with higher cognitive abilities made more consistent choices and *not* less risk-averse choices. In line with this notion and in contrast to what dual system theories would predict, is research showing that sometimes even stronger behavioral biases exist in people with higher cognitive capacities (Corbin, McElroy, Black, 2010).

At this point, we have two competing hypotheses about the link between WMC and risky decisions. Either smarter people are less risk averse because of a greater reliance on an expected-value-calculating System 2. Or, smarter people are less noisy and, because of the

design of the studies that have been run so far, these more consistent decisions have been mistaken for less risk averse decisions.

Decision by sampling predicts cognitive load will reduce sensitivity to the distribution of attribute values

We can also use decision by sampling (DbS, Stewart, 2009; Stewart, Chater & Brown, 2006) to make predictions about the effect of cognitive load on risky decision making. DbS postulates a major role for working memory in the decision making process. In DbS, working memory is required for all three of its constituent mechanisms: sampling recently encountered attribute values, pairwise comparisons between the attribute values, and an accumulation process which counts the number of favorable comparisons. As we explain below, DbS predicts that, when working memory affects these three processes, people's risk propensity should be less sensitive to the broader context in which their decisions are made.

Maintenance and retrieval of items are the two main functions of working memory (Unsworth & Heitz, 2007). Thus, if participants are under high cognitive load when they make a decision, the attribute values they are considering could either be pushed out by the load items, the load items could interfere with the relevant attributes, or the accumulation process could be impaired. Consequently, cognitive load should diminish the context sensitivity demonstrated in a recent study by Stewart, Reimers and Harris (2015). In that study, participants' utility functions were found to be strongly affected by the distributions of prizes, probabilities, and times on offer during the course of the experiment. More specifically, when participants were presented with a positively skewed distribution of prizes, the utility function estimated from people's choices was concave and participants' choices were risk-averse. In contrast, when participants were presented with a negatively skewed or uniform distribution of prizes, the estimated utility function was convex or linear, respectively, with participants appearing to make risk-seeking or risk-neutral decisions. As in Chapter 2, we call the sensitivity of the preference functions to the attribute distributions, the

SRH effect. Will the difference between the utility functions for differently distributed prizes remain for choices under high cognitive load? This is a test of the specific mechanism in DbS, by which the distribution of attribute values influences choices via a working memory mechanism.

Thus, DbS adds a third hypothesis to the set: Reduced working memory capacity should impact the proposed cognitive processes. If risk aversion is the result of the interaction between cognitive processing and the distribution of attribute values in the environment, reduced working memory capacity should reduce this interaction. This third hypothesis was the motivation for the design of the experiments in this paper. However, our data allowed us to test the first two hypotheses as well, so we did.

Experimental Programme

In the following studies we crossed two factors: the distribution of attribute values in the choice set with the working memory load imposed within participants whilst the choices were made. Decision by sampling predicts that the effect of the distribution of attribute values should be reduced with high cognitive load. The dual system account predicts that decision making should be more risk averse under cognitive load, irrespective of the distribution of attribute values. The increased-noise account predicts that decision making should be less consistent or deterministic under cognitive load, irrespective of the distribution of attribute values. Thus the aim of manipulating cognitive load was to be able to answer three questions: (a) How robust are rank effects in the presence of cognitive load within participants? (b) Will participants make more irrational (i.e., more risk averse) choices under cognitive load? (c) Will participants make less consistent responses under cognitive load?

We are presenting three versions of a similar experimental set-up where participants make consecutive risky choices partly under no or low cognitive load, partly under high cognitive load, to investigate the effect of high cognitive load on the context sensitivity of valuation. From a theoretical point of view, the studies could help identify boundary

conditions, namely whether people's choices might follow the DbS proposed mechanism only when enough cognitive resources are available to the decision maker. Ultimately, establishing whether cognitive abilities influence risk preferences matters, given that recommendations for policy makers would differ if people were risk averse due to their low cognitive abilities, or if their choices were just random because low cognitive abilities lead to noisier decisions.

Experiment 1A

Method

Participants. 54 Warwick psychology undergraduates participated for course credit. After the experiment, a random gamble was played out so that participants could win up to £5 depending on the choice that they made in the experiment.

Design. Following Stewart et al. (2015) we created 30 gambles with a positively skewed amounts distribution where participants encountered £10, £20, £50, £100, £200, and £500 throughout the experiment. For the uniform amounts distribution, we created 25 gambles where participants encountered £100, £200, £300, £400, and £500 throughout the experiment. The probabilities were 20%, 40%, 60%, 80% and 100% percent in both conditions. We crossed all gambles and dropped the pairs where one gamble stochastically dominated the other so that participants in the uniform condition made choices out of a possible set of 100 pairs and participants in the positive skew condition made choices out of a possible 150 pairs.

Procedure. Participants were instructed that they were going to make 75 choices, they could take as much time as they liked and that depending on their choices they could win a bonus. 75 pairs out of the 100 and 150 possible pairs were randomly selected. For each trial, participants were faced with either a string of two (Low-load) or seven random letters (High-load), which they were told to memorize and recall when asked. Whether participants were under low cognitive load or high cognitive load per trial was random too. They could take as much time as they liked to remember the string, before proceeding to the pair of gambles to

choose from. Each gamble was presented on the screen as a button and participants indicated which gamble they preferred by clicking on it. When they made a decision, the recall screen appeared with a textbox where participants were asked to fill in the letter they had memorized beforehand in the same order. The experiment was over after all choices were made, one gamble was played out and participants received their compensation.

Results and discussion

Manipulation check. We first tested whether participants performed better during recall, i.e. recalled more strings correctly, when under low than under high cognitive load, which they did, *Mean difference (MD)* = .32 95%CI [.26, .38]. This shows that participants indeed were under load when trying to memorize the string and simultaneously making a decision and that remembering seven letters was harder than remembering two. We also checked whether there are differences in recall between the two distributions and cognitive load; there were no interaction effects, MD_{Low} = .01 95%CI [-.01, .04] and MD_{High} = .06 95%CI [-.06, .18].

Context sensitivity. Stewart et al. (2015) reports the details of the estimating procedure for the value of each amount for the two conditions (Positive-skew and Uniform) and for bootstrapping confidence intervals. We used mixed effects logistic regressions separately for the two distributions, using cognitive load as a within-subjects variable to estimate the value for trials in the low-load condition and trials in the high-load condition. Figure 3.1 depicts the value function for the positive-skew and the uniform conditions, the left side shows low-load condition trials, the right side shows high-load condition trials. Both patterns replicate previous results where a positively skewed distribution yielded a more concave and the uniform distribution a linear value function. The effect sizes of the effect of distribution, estimated using the differences between the estimates of £200, are very similar between Low-load and High-load, mean difference for the Low-load (MD_{Low}) = .34 95% CI [.08, .59] and MD_{High} = .26 95% CI [.03, .49]. The pattern for Low-load was predicted by DbS

and replicates previous findings. The pattern in High-load largely remains, contrary to predictions of DbS.

Risk aversion. To test whether people were more risk averse under high load (i.e., whether the functions are more concave under high cognitive load), we chose to compare the estimate of the amount of £200 within the corresponding distribution between high and low cognitive load. Cognitive load was a within-subjects variable in these experiments and the model estimates take that into account: In the positive-skew condition the estimate in the high-load condition is slightly lower than in the low-load condition, and only slightly higher in the uniform condition, $MD_{Positiove\_200}$= .04 95% CI [-0.07, .15] and $MD_{Uniform\_200}$= -.02 95% CI [-.20, .15]. More cognitive load does not lead to more risk-aversion; The estimates are not reliably higher in the high-load condition, which would make the function more concave–indicating more risk aversion–in the high-load conditions than in the low-load conditions.

Noise. Using γ, the determinism parameter, which captures the degree to which participants choose consistently across trials, we can estimate how noisy participants' choices were (see Stewart, Reimers & Harris, 2015 for details). According to Andersson et al. (2013), high cognitive load should lead to less consistent choices. We do find higher estimates for γ in the low-load than the high-load conditions, suggesting more noisy responding in the high-load conditions. Yet, our best estimate reveals only small differences between γs for low-load and high-load trials, $MD_{Positiove\_\gamma}$ = -0.08 95%CI [-0.17, 0.01] and $MD_{Uniform\_\gamma}$ = -0.06 95%CI [-0.19, 0.08].

Cognitive load does not seem to affect risk aversion but it does affect consistency a little, according to our data. To gain further insight whether cognitive load indeed has no effect on context sensitivity, we have replicated this experiment online with a larger sample.

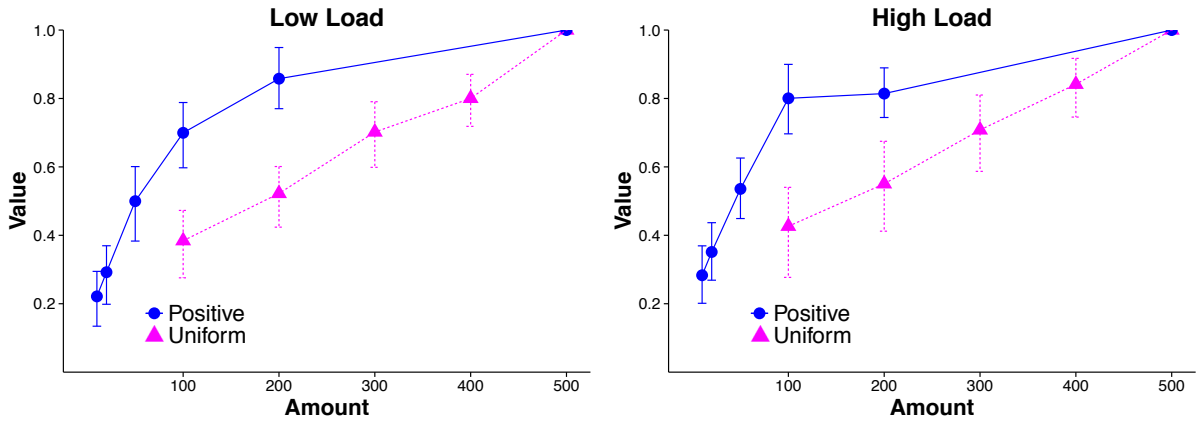

Figure 3.1. Utility functions for Experiment 1A. Error bars are 95% confidence intervals.

Experiment 1B

Method

Experiment 1B replicates Experiment 1A with a larger sample using Amazon Mechanical Turk. 145 people participated for $1.80. The only other difference between 1A and 1B was that here participants were not incentivized beyond us urging them that reporting their true preferences and trying their best during the memorization and recall task was very important for the task.

Results and discussion

Manipulation check. Participants again were far more accurate during the recall task when under low than under high cognitive load, *MD* = .39 95%CI [.34, .43]. The difference here is a little bigger than in Experiment 1A, because accuracy was a little higher during the Low-Load trials and a little lower during the High-Load trials as compared to Experiment 1A. However, the differences are within a similar range. There are no interaction effects of load and distribution on the proportion of accurate recalls, MD_{low} = .01 95%CI [.00, .02] and MD_{high} = .06 95%CI [-.03, .15].

Context sensitivity. Figure 3.2 depicts the value function for the positive-skew and the uniform condition. As above the left side shows estimates for low-load condition trials, the right side shows estimates for high-load condition trials. In the low-load condition, the

positive-skew condition compared to the uniform condition resulted in a more concave utility function. However, the difference is much smaller than in Experiment 1A, $MD_{low=}$ .16 95% CI [-.09, .40]. In the high-load condition, the differences between value functions of the positive-skew and uniform conditions are even smaller than in the low-load condition and almost collapse, $MD_{high=}$ .05 95% CI [-.23, .34]. This indicates that cognitive load might slightly desensitize participants' choices to the distribution they experience.

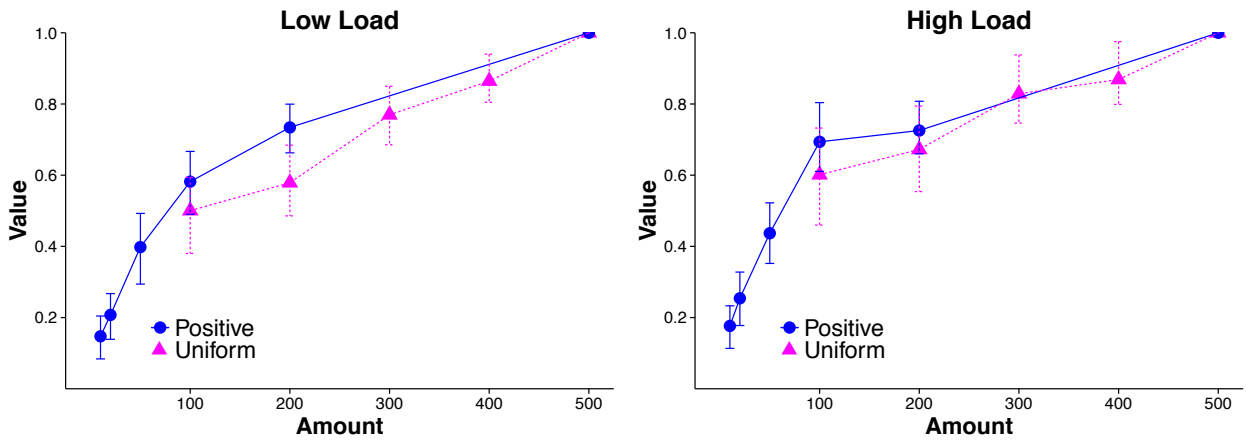

Figure 3.2. Utility functions for Experiment 1B. Error bars are 95% confidence intervals.

Risk aversion. We also see almost no change in risk aversion with cognitive load. When comparing the high-load estimates with the low-load estimates for the positive-skew and the uniform condition separately, cognitive load has no effect, $MD_{Positive\_200}$= .01 95% CI[-.09, .11] and $MD_{Uniform\_200}$= -.09 95%CI [-.24, .06].

Noise. When comparing γ for the high-load against the low-load trials, γ in the high-load conditions are estimated to be a little lower than in the low-load conditions, again showing a small effect on noise within participants, $MD_{Positive\_\gamma}$ = -0.07 95%CI [-0.08, -0.07] and $MD_{Uniform\_\gamma}$ = -0.12 95%CI [-0.25, -0.00].

From previous experiments, we know that participants are sensitive to the distribution manipulation when making sequential choices. However, in the 1B low-load conditions the differences between the positive-skew and the uniform conditions were much smaller than in Experiment 1A and other previous experiments. Failing to find a contrast between high-load

and low-load trials in both 1A and 1B, led us to slightly change the experimental design in Experiment 2. In this experiment too we have consistent findings across conditions considering choice consistency, which we will also test for in the next experiment.

Experiment 2

We made the following changes to the design: First, we decided on a block design. Instead of randomly selecting low load and high load trials, all participants first made choices without load and then proceeded with the high-load trials for the last choices. Second, to be able to purely estimate the effect of cognitive load, for the first part of the experiment we decided to completely omit the low cognitive load part of the trials. Both these changes were supposed to ensure that participants' working memories actually contained the attributes we wanted to manipulate and to reduce interference. We feared that in the first two experiments the ongoing load, low or high on every trial, was sufficient to prevent participants from learning the distribution of attribute values as well as we might expect them to, leading to the findings in 1B.

Method

Participants. Because we failed to detect differences between the low-load and the high-load conditions in Experiments 1A and 1B, we ran Experiment 2 with 492 participants via Amazon Turk. Each received $1.80 for taking part. Based on a power analysis and the effect sizes measured in SRH-type experiments, we estimated that sample size needed to be tripled to have a 99% chance of detecting the SRH effect, i.e. differences between the distribution manipulation conditions.

Procedure. For the first 40 trials participants only made choices between gambles where the memorization and recall tasks were completely omitted. After the no-load trials participants proceeded to the 20 remaining high-load trials with the same procedure as in Experiment 1A and 1B: Memorization was followed by the choice between two gambles,

which was followed by recalling the seven letter string. As before, participants were told that they could take as much time as they liked to complete the task.

Results and discussion

Manipulation check. Because there were no low-load condition trials, here we only checked whether recall was equally poor for both distribution conditions, *MD* = .03 95%CI [-.02, .08], making sure that remembering a seven-letter string was hard enough to induce cognitive load independently of the distribution of amounts participants experienced.

Context sensitivity. The estimated value functions are depicted in Figure 3.3. The original distribution manipulation effect was replicated under no cognitive load: In the positive-skew condition, participants' utility functions were more concave than in the uniform condition, $MD_{No\_Load=}$ .15 95% CI [.05, .24]. Under high cognitive load, the SRH effect is of similar size as under no load, but with a less reliable estimate, $MD_{High\_Load=}$ .11 95% CI [-.06, .27].

Risk aversion. We also see, that cognitive load has no effect on risk aversion within participants in the positive-skew condition, $MD_{Positive=}$ .00 95% CI [-.06, .06] and very little effect in the uniform condition, $MD_{Uniform=}$ -.04 95% CI [-.09, .01].

Noise. The wider confidence intervals on the estimates in the high-load condition indicates that choices might be noisier under high cognitive load compared to no cognitive load. Indeed, when looking at the γs, both the positive-skew condition and the uniform condition–though with wide confidence intervals here–yield lower γs in the high-load condition than the no-load condition, $MD_{Positiove\_\gamma}$ = -0.58 95%CI [-0.87, -0.28], and $MD_{Uniform\_\gamma}$ = -0.15 95%CI [-0.76, 0.46]. The findings across experiments are consistent with Andersson et al.'s hypothesis about the negative effect of load on choice consistency.

Across all three experiments, we found that the effect of the distribution manipulation replicated. There is very little evidence that cognitive load had an effect on risk aversion, but we consistently find that cognitive load decreases choice consistency. To get an overview

across experiments, and in order to estimate the effect of cognitive load on the SRH effect, on risk aversion and choice consistency, we conducted three meta-analyses.

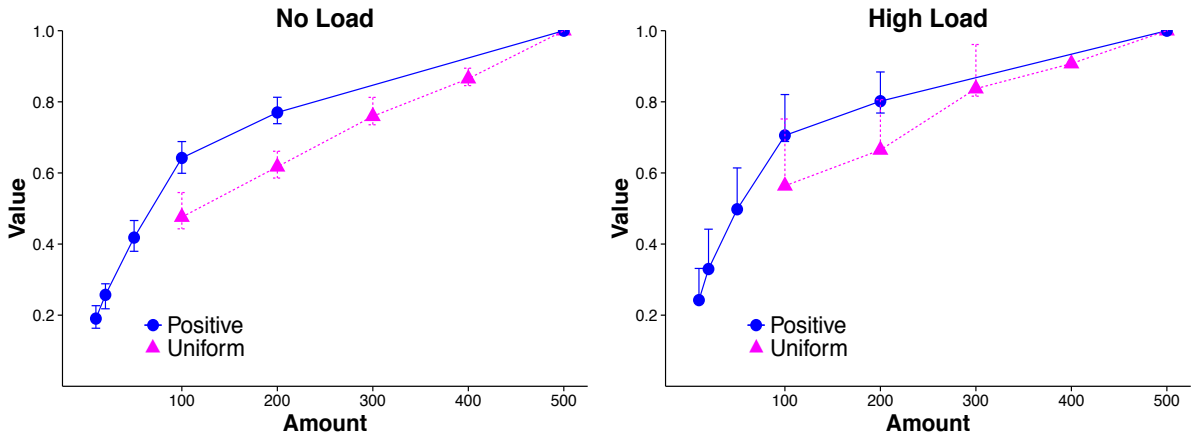

Figure 3.3. Utility functions for Experiment 2. Error bars are 95% confidence intervals (CIs are asymmetrical in the high-load conditions and are hidden under the dots and triangles).

Meta-analysis

Above we have described three experiments, with a mostly consistent pattern of results: The original distribution manipulation effect replicated four times in Experiment 1A and Experiment 2, but did not replicate in neither the low-load nor the high-load conditions in Experiment 1B. Given all previous findings (see Chapter 2), we have reason to believe that this possibly occurred due to random variation. To conclusively answer our main question, we combined the evidence and tested whether on aggregate the effect size of the distribution manipulation is reduced by high cognitive load as compared to no/low cognitive load. We also add estimates of the effect sizes of noise and the risk aversion hypotheses. All analyses will yield a conservative estimate of the difference between high and no/low cognitive load, because they do not take into account the within-subjects nature of our data (see Dunlop, Cortina, Vaslow, & Burke, 1996 for reasons why this is the appropriate procedure). The models used to estimate the parameters do however take the within-subjects design into account.

The SRH effect remains under high cognitive load

Because the experiments are not exact replications of each other, we implemented a random effects meta-analysis using the metafor package for R (Viechtbauer, 2010). Following our previous meta-analysis, we used the differences in the estimate of the subjective value of £200 (or $200 for Experiment 1B and 2) of the positive-skew and the uniform condition, as our measure of the effect of the distribution of attribute values. We separated the effects by adding cognitive load as moderating variable. Decision by Sampling predicts a smaller difference between the Positive-skew and the Uniform condition under high cognitive load than under low/no cognitive load, because the WM component, presumably involved in the choice process, should be impaired by cognitive load.

Figure 3.4 shows effect size estimates for Experiments 1A, 1B, and 2, separately for no/low-load and high-load conditions. The estimate of the difference between the subjective values of £200 in the positive-skew and uniform conditions for the no/low-load conditions is 0.17 95% CI [0.10, 0.23] on a scale where the value for £0 is zero and the value for £500 (or $500) is set to be 1. The estimate of the difference in the high-load conditions is .16 95% CI [0.03, 0.29]. The estimates strongly resemble one another, and the confidence intervals of the High-load estimate completely encompasses the No/Low-load estimate. The random-effects model estimates the difference between risk aversion on the no/low-load and the high-load condition trials to be nearly zero, $beta_{Load}$ = -0.01 95% CI [-0.15, 0.14].

The meta-analysis has helped to identify that even on aggregate cognitive load does nothing to reduce the SRH effect. In addition, we can compare this overall estimate with the overall estimate in Chapter 2, where we found the difference between the subjective values of 200 to be $MD_{Overall\_Chapter2}$ = 0.17 95% CI [0.13, 0.22], which is almost identical to the overall estimate in this chapter, $MD_{Overall}$ = 0.16 95% CI [0.10, 0.23].

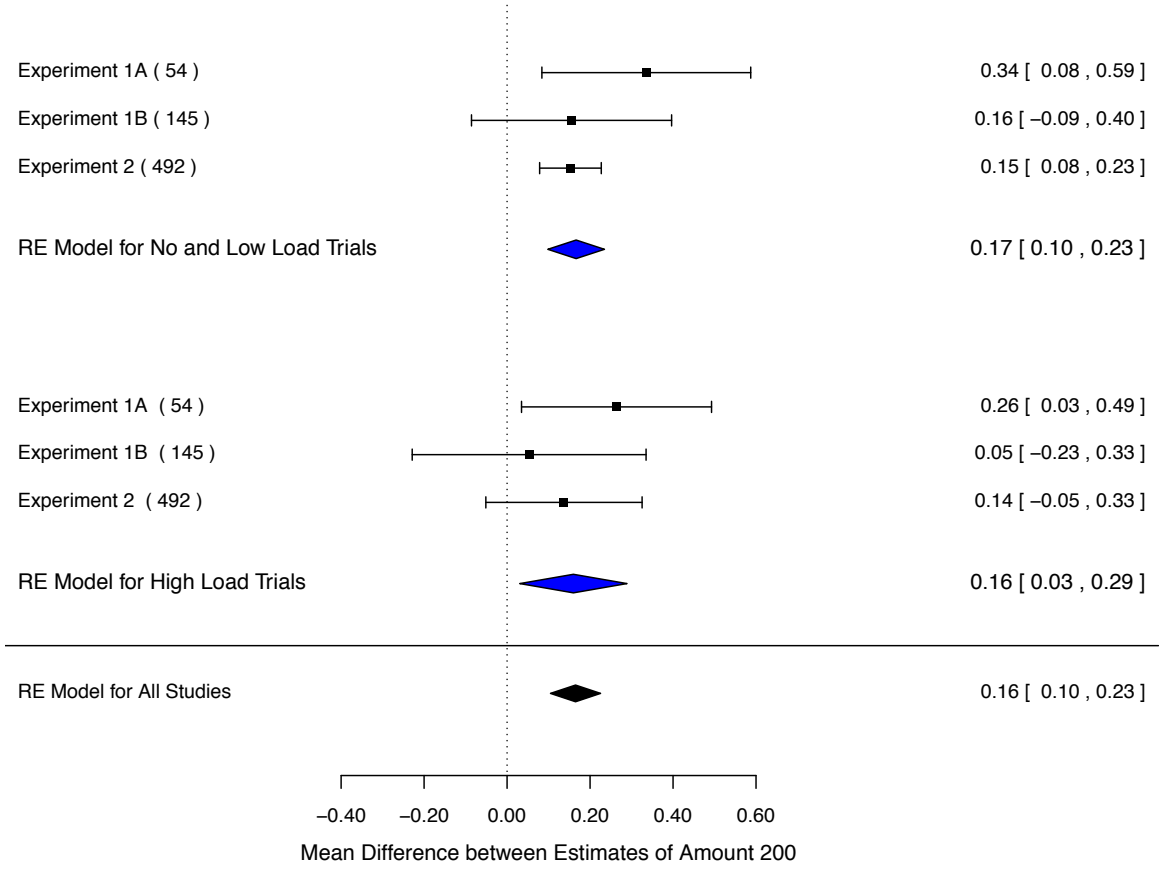

Figure 3.4. Mean differences between the estimate for £200 (and $200 respectively) in the positive-skew and the uniform conditions with 95% confidence intervals. The estimates are shown as a function of the presence and the amount of cognitive load.

Risk aversion does not increase under high cognitive load

To estimate, whether cognitive load increases risk aversion, as above we used the differences in the estimate of the subjective value of £200 (or $200 for Experiment 1B and 2) for low cognitive load trials and compared them with estimates for high cognitive load trials. We separated the effect between positive-skew and the uniform condition. Figure 3.5 shows effect sizes for risk aversion estimates in Experiments 1A, 1B, and 2, separately for positive-skew and uniform conditions. The estimate for the difference in subjective values of £200 between the no/low-load and high-load conditions overall is -0.01 95% CI [-0.05, 0.02] on a scale where the value for £0 is zero and the value for £500 (or $500) is set to be 1. The estimate of the difference in the uniform conditions is -0.04 95% CI [-0.09, 0.01] and in the

positive conditions it is 0.01 95% CI [-0.04, 0.06]. The random-effects model estimates the difference between risk aversion in the positive-skew and the uniform conditions to be very small, $beta_{Distribution\_RiskAversion}$ = 0.05 95% CI [-0.01, 0.11]. In one experiment we find a decrease in risk aversion with high cognitive as compared to no/low cognitive load, in half we find no effect and in two conditions there indeed was an increase in risk aversion under high cognitive load. However, contrary to what is implied by dual-system theories, the effect of cognitive load on risk aversion given our data is basically zero.

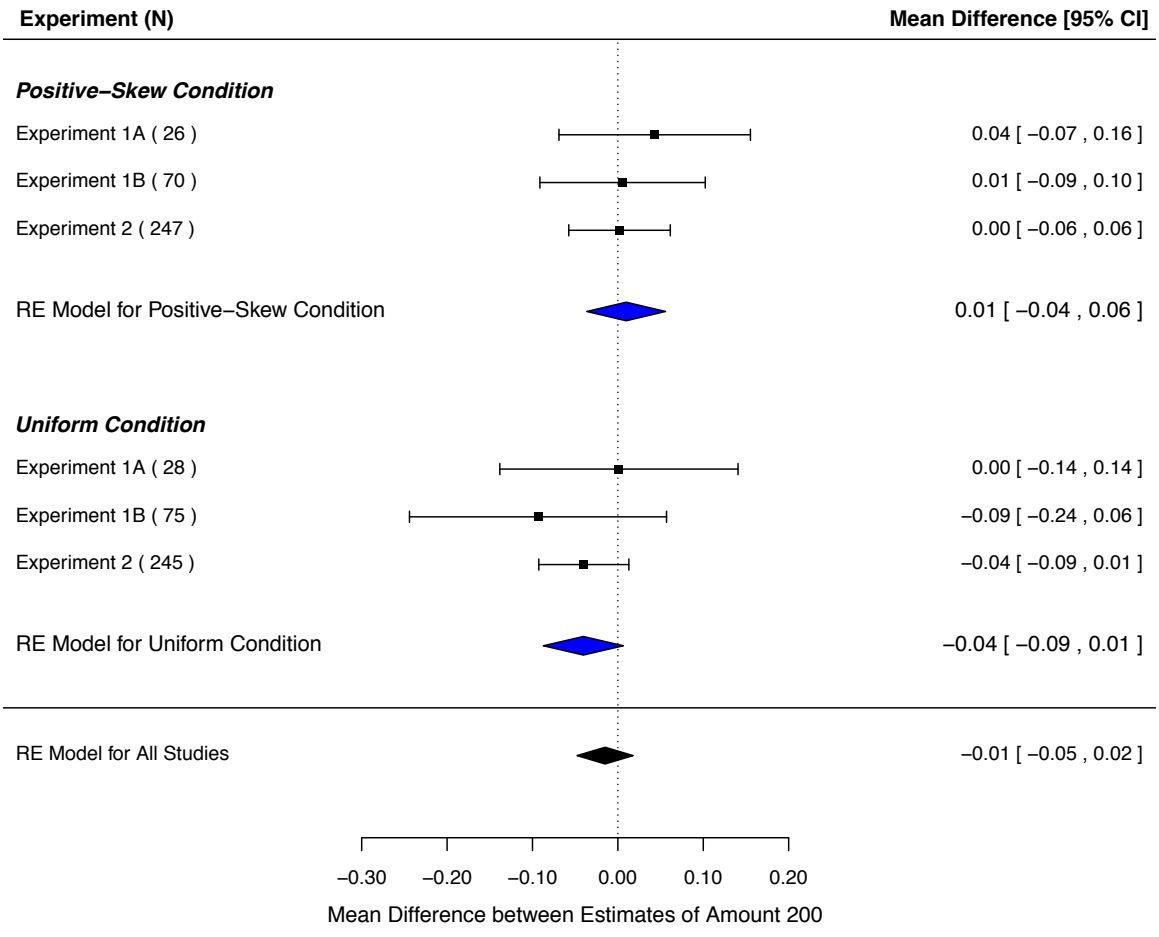

Figure 3.5. Mean Differences between the estimate for £200 (and $200 respectively) in the no/low load and the high load conditions with 95% confidence intervals, shown as a function of distribution (Positive-skew or Uniform)

Cognitive load decreases consistent choice behavior a little

We used the differences in γs between the no/low load and high load condition to

estimate the consistency of participants' choices. In all our experiments we found lower γ in the high-load condition than the no/low-load condition. The overall effect size is -0.08 95% CI [-0.08, -0.07] with an effect of -0.21 95% CI [-0.50, 0.08] in the positive-skew condition and -0.15 95% CI [-0.76, 0.46] in the uniform condition. Figure 3.6 shows all effect sizes separately for the positive-skew and uniform conditions. The random-effects model estimates the difference between γs in the positive-skew and the uniform conditions to be very small, $beta_{Distribution\_Consistency}$ = -0.08 95% CI [-0.37, 0.21]. As in Anderson et al. (2013), high cognitive load does decrease choice consistency. However, on a scale where the γs are estimated to range from two to four, the overall effect we observe here is very small.

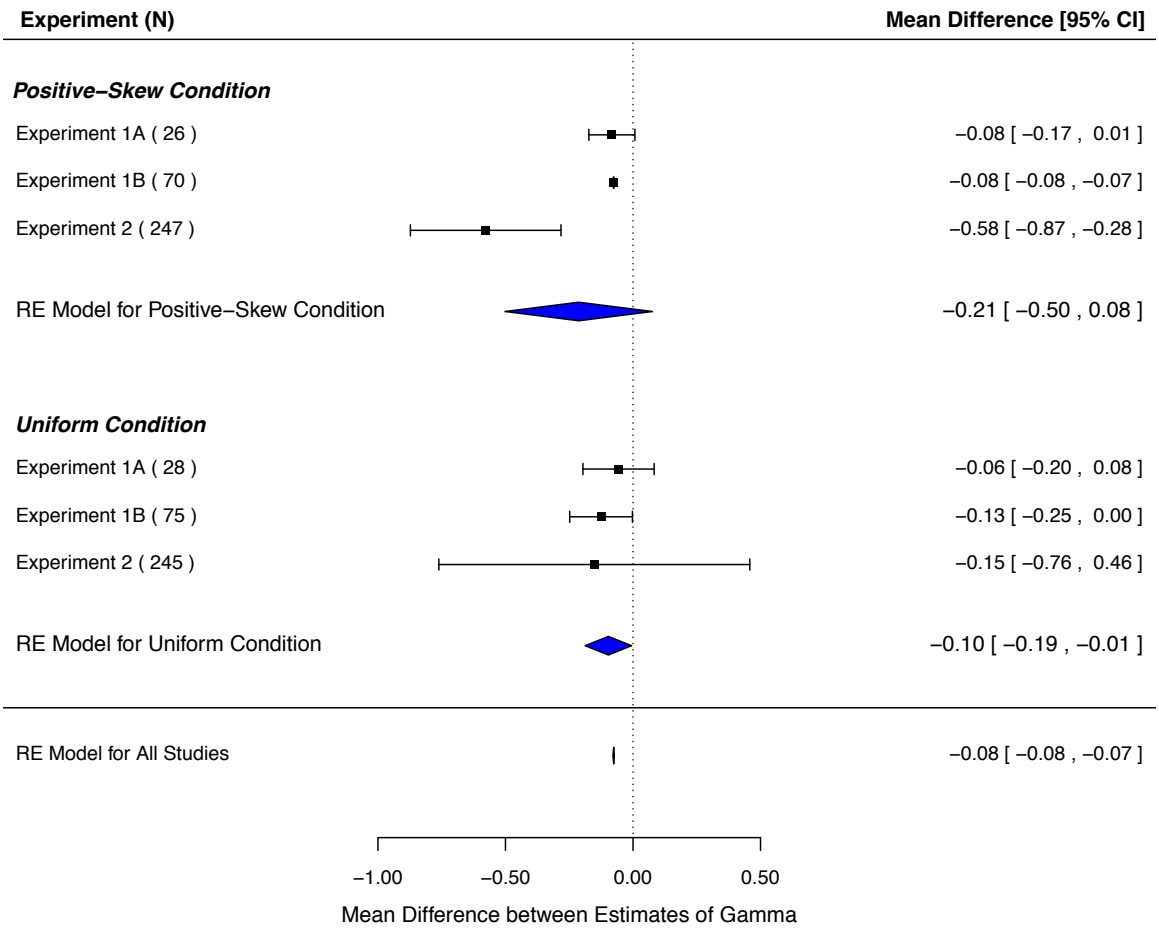

Figure 3.6. Mean Differences between estimates of γ in the no/low load and the high load conditions with 95% confidence intervals, shown as a function of distribution (positive-skew

or uniform condition). Γs are consistently lower in the high-load than the no/low-load conditions.

General Discussion

By asking people to make choices between two gambles under no/low and high cognitive load, our study primarily investigates whether context sensitivity is reduced in the presence of high as opposed to no/low cognitive load. Previous results looking into the impact of cognitive load on risky choice are inconclusive (see Deck & Jahedi, 2015 for a recent review): Some find an increase in risk-averse choices (Dohmen et al., 2007; Oechssler, Roider & Schmitz, 2009), others find a decrease in choice consistency with lower cognitive abilities or under high cognitive load (Andersson et al., 2013; Franco-Watkins, Rickard & Pashler, 2015). Looking at our data, we find no evidence that high cognitive load leads to more risk averse choices (High-load utility functions are not more concave than No/Low-load utility functions). We do find however, that high cognitive load leads to noisier choice behavior, although the effect is very small. Although participants' choices are noisier under high cognitive load, they are still sensible, given that all utility functions increased monotonically.

To understand how our choices are impacted by cognitive load, Andersson et al. (2013) recommend using models that distinguish parameters for preferences from parameters for somebody's tendency to make mistakes. They also recommend measuring cognitive abilities in order to be able to factor it in when eliciting choice preferences. From a theoretical point of view this makes a lot of sense, if enough evidence corroborating their finding is accumulated and if we started using a more diverse sample as opposed to student samples. From a practical point of view, it might not pay off if the sample is too homogenous.

What in other studies might be mistaken for risk aversion, might as well be noise (and we should find out if this is the case). Our model does allow us to distinguish between the two. Probably also of importance is that our model also allows us is to estimate the effect size of cognitive load on choice consistency: We only find a small effect size and Andersson et al.

do too (correlation coefficient of 0.09 and 0.1 between cognitive abilities and noise). Even if the effect size could be influenced by our cognitive load manipulation and no control of cognitive capacities, it needs to be considered when evaluation these findings. Still, they serve as another lesson in how wrong models can lead to spurious results and wrong conclusions.

Besides Bruza, Welsh and Navarro (2008), where they investigated framing effects controlling for working memory on the basis of DbS, to the best of our knowledge, our experiments are the first direct test of the involvement of working memory considering DbS predictions. We found no effect of working memory in moderating the association between the distribution of attribute values in the choice set and the revealed utility functions. One possible reason for this is our choice of cognitive load task. We deliberately chose to have letter strings rather than digit strings, because we did not want people to be comparing the digit strings to the attribute values. But perhaps there is a separation of the numerical attribute values experienced from the letters that are to be remembered. Thus, should working memory's content facet consist of a separate verbal and numerical category, the numbers and letters could be stored and retrieved separately (see Oberauer, Süss, Schulze, Wilhem & Wittmann, 2000) and the risk of interference would be diminished.

Our experiments ignore response times. The additional constraint on capacity by high cognitive load could be compensated with taking more time to make a choice. This could be tested by testing whether participants take longer to respond under high rather than low cognitive load. An additional factor that we have not accounted for, which might have led to no impact of high cognitive load, is participants' operation span. However, given that we tested around 700 people, it is unlikely that a small number of participants with very high cognitive capacities have distorted the data. Still, it would have been advantageous to have controlled for operation span, considering that Deck and Jahedi (2015) in their study found that participants whose answers were most negatively affected by high cognitive load (i.e., participant with low operation span), were also the participants who under cognitive load

made more risk averse choices.

Another explanation of the SRH effect could be offered by Parducci's range-frequency theory (RFT; 1965). In RFT the judgment of an attribute is a compromise between the rank of an attribute within a set of other present attributes and its position within the range of the set of the same attributes. RFT is not formulated as a process model, but assumes that any kind of judgment or evaluation of some sort of magnitude is a weighted mix of positions within the rank and range of stimuli. While DbS hypothesizes a rank effect arising from binary ordinal comparisons, RTF does not postulate a mechanism. Instead the argument is one of optimality: the rank principle leads to a reduction in reconstruction error if representation capacity is limited. Anyway, it could be that the rank effects in Stewart et al.'s (2015) experiment arise through some non-working-memory mechanism that reflects this rank principle. For example, if participants stretch their internal scale that represents the sums of money in the real world so that there is greater resolution where amounts are more frequent, to mirror the distribution of the amounts presented here, we would need no specific working memory part to observe these results, even under high cognitive load.

Across different experiments, the rank hypothesis holds. In previous work, Stewart, Chater, and Brown (2006) really just highlighted a coincidence: With the shape of the distribution of attribute values in the world, and the mechanism proposed by DbS, the authors observed that utility, weighting, and discounting functions had the same qualitative shape as described in the literature. Ungemach, Stewart, and Reimers (2008) and Stewart (2009) reported some experimental evidence to support this link. By manipulating the distribution of attribute values that people recently experienced, they were able to change people's decisions in a way DbS predicted. Stewart, Reimers, and Harris (2015) extended this result, showing the direct manipulations of the skew of the distribution of attribute values indeed alters decisions, with revealed utility functions changing just as predicted by DbS. However, in Chapter 2 we showed that it cannot all be DbS, but that the reason the functional forms take the shapes they

do, might as well be completely independent of the samples of attribute values in people's working memory. We estimated that only about half of this effect should be attributed to the DbS mechanism, with at least one half as yet unexplained. The fact that cognitive load does nothing to diminish the effect size under our experimental circumstances, does not help in finding the mechanism leading to about half of the apparent rank effects found in our studies. All previous work on DbS did hypothesize that working memory was a key component, responsible for the effect of the distribution on choice. With the design used for the experiments in this chapter, we could not provide any evidence that differences in working memory capacity lead to changes in the SRH effect, leaving us with more questions about its exact role in risky choice.

4 Stewart & Reimers (2008): More evidence for the rank hypothesis in judgment and choice

We know the SRH effect is robust across different experiments. However, we also know from our within-subjects design experiments from Chapter 2 (Alempaki, Canic et al., 2016), in which participants made choices across both sets of amounts, that not the whole effect, and perhaps not any of the effect should be attributed to decision by sampling (DbS, Stewart, Chater, & Brown, 2016). We know that at least to some degree, the effect is caused by systematic biases in the estimation procedure, which vary as a function of the attribute values presented (Stewart, Canic, & Mullett, in prep). One might wonder, whether these results entirely question the decision by sampling model. We will show that DbS predicts many findings that are unaffected by these estimation problems. First, we will reintroduce decision by sampling and older theories that motivated the model. Then we will review behavioral and neurophysiological evidence that is in line with the rank hypothesis and unaffected by the estimation problems causing reported in Chapter 2. Lastly, we will present yet unpublished data by Stewart and Reimers (2008) and will estimate the rank effects in these experiments.

That context influences judgment has been established well before Stewart, Chater, and Brown proposed DbS, especially in psychophysical models dealing with perceptual judgments. In his adaptation level theory (ALT), Helson (1964) assumed that judgment about magnitudes, like the intensity of sounds or the brightness of lights, are made relative to a reference point–the adaptation level–which was often taken as the attributes' average. Parducci (1959) tested ALT predictions for the effect of presentation order, i.e. whether the stimuli were presented in increasing, decreasing order or whether participants saw all stimuli before they started judging their magnitude by assigning them into categories. He also tested how the number of judgment categories influenced judgment. Parducci found that not only the order of judgment but also the order of presentation, biased judgment towards the first presented (not consistent with ALT) and the first judged (consistent with ALT). In addition,

the number of judgment categories also strongly affected people's judgments. Specifically, Parducci found that if the number of categories was increased or the number of stimuli was decreased, such that there were a lot of categories relative to the number of stimuli, the effect of order of judgment was strongly reduced. Other experiments involved looking at effects of stimulus extremes, the measure of central tendency (mean, midpoint of the judgment scale, or median) or other effects of the distribution of stimuli (Parducci, 1963). Varying the mean had no substantial effect on AL. However, varying the midpoint of the stimulus extremes and the median independently of each other and holding the mean constant, had an impact on AL (Parducci et al., 1960): The stimuli around the middle of a distribution are judged to be lower when the midpoint between two stimulus extremes is high instead of low and when the median is low instead of high. Parducci's (1963) key insight was that while adaptation level theory captures the finding that people evaluate magnitudes against the mean of the distribution from which they come, it foregoes the important property that people are also sensitive to the variance of the distribution of values. For example, being £10 above the mean seems more significant when all of the other values are within +-£5 of the mean rather than +-£50 of the mean. Given his results, Parducci concluded that ALT had to be adjusted, because it does not describe what principles are actually critical for judgment.

Parducci (1965) assumes participants' judgments are subject to two tendencies: The first tendency is to divide the range into subranges, where each category of judgment is applied to a subrange. The second tendency is to use a category proportionately to the number of times one makes a judgment. Assuming these two tendencies, Parducci (1965) formulated range-frequency theory (RFT) to account for people's sensitivity to the spread of magnitudes, which are to be evaluated, and their relative frequencies. According to RFT, judgment is a compromise between a stimulus' position within the range of the contextual stimuli (the range principle) and the stimulus rank among the contextual stimuli (the frequency principle).

When Stewart, Chater and Brown (2006) proposed the decision by sampling model, they also referred to RFT. The process described in DbS reflects RFT's frequency principle, which makes the two theories closely related. Despite arriving at the same principle, the models are distinct in that DbS takes its inspiration from established aspects of judgment and choice concerning domain-general simple cognitive tools: ordinal comparisons, sampling and frequency accumulation. In DbS, an attribute is evaluated through pairwise comparisons between a current attribute and attributes from memory, and the attributes are evaluated according to their rank position within this memory sample. The fact that it's easy for people to establish whether one stimulus is louder, brighter, or longer than the other, but it's hard for people to establish their absolute magnitude (Laming, 1997), makes the ordinal comparison component in DbS a plausible one. Other authors in their theories have also considered DbS' sampling from memory and subsequent comparisons and their accumulation as important aspects. In norm theory, Kahneman and Miller (1986) proposed that the evaluation of a current outcome depends on alternatives that are implied by a norm. Here, norms are constructed on the fly, using the specific attributes available at the moment of evaluation. In decision field theory (Busemeyer & Townsend, 1993), evidence is accumulated sequentially and the decision is made when a threshold is hit. In MINERVA-DM, hypotheses are supported depending on the similarity to traces from memory (Dougherty, Gettys, & Ogden, 1999). In query theory, people make a choice according to the accumulated answers they give to internally proposed and ordered queries, whose order depends on the relationship of the decision maker to the problem at hand (Weber et al., 2007). Relativity is also an aspect of fuzzy-trace theory (Reyna, 2008), where people make choices using fuzzy representations, which could be relative.

To demonstrate the relativity of option values, Ungemach, Stewart, and Reimers (2011) asked participants to choose between two gambles, just as they came out of a supermarket: choose Gamble A with £1.50 and a low probability or Gamble B with £0.50 and

a high probability. They found that if participants purchased goods in-between the amounts of £0.50 and £1.50 (inside-range, they preferred Gamble A. However, if participants purchased goods outside this range, either smaller than £0.50 or larger than £1.50 (outside-range), then participants' preferred Gamble was B. This preference reversal is exactly what DbS predicted: Participants chose as if £1.50 was very different from £0.50 (big difference in rank), because many other amounts laid in-between, whereas in the latter case all prices outside this range made £1.50 seem very similar to £0.50 (very small difference in rank). There are several other experimental demonstrations of context effects in preference, where either choices sets or the response scale was manipulated, that all fit into a DbS account. Table 4.1 summarizes the key features of the experiments presented below.

Behavioral Evidence

Over the course of several years, many studies have either investigated context effects in risky choice and judgment or context effects were a by-product of these investigations. We have divided the studies into three groups of behavioral studies where either the choice set was manipulated, studies where the response scale was manipulated, and others where the valuation background was manipulated. The studies' dependent variables were choices, using gambles, or judgments like attractiveness ratings, buying prices or certainty equivalents. Within those sections, most studies looked at range effects, and only a few manipulated the skew of the distribution. Both these manipulations can lead to different ranks of specific target values, which in turn make a DbS account testable. We will present studies using manipulations of the background context first, move on to choice set manipulations and lastly present studies with response scale manipulations.

Table 4.1

Key features of experiments investigating context effects in judgment and choice.

Measure	Author	Manipulation of	Response Mode	Distribution Manipulation
Behavioral	Adaval & Monroe (2002)	Background context	Price rating	Range
	Beauchamp et al. (2012)	Response scale	Choice	Skew
	Benartzi & Thaler (2001)	Response scale	Choice	Range
	Birnbaum (1992)	Response scale	CE	Skew
	Bohm et al. (1997)	Response scale	Reservation price	Range
	Ert & Erev (1997)	Choice set	Choice	Range
	Haggag & Paci (2014)	Response scale	Taxi tipping	Range
	Janiszewski & (1999)	Background context	Price rating	Range
	Mazar et al. (2013)	Response scale	Reservation price	Skew
	Mellers et al. (1992)	Choice set	Prospect rating and buying price	Skew

	Rigoli, Rutledge et al. (2016)	Choice set	Choice	Range
	Stewart et al. (2003)	Response scale and choice set	CE and choice	Range
	Stewart & Reimers (2008)	Choice set	Rating and Choice	Skew and Range
	Ungemach et al. (2011)	Background context	Choice	Range
	Vlaev et al. (2007)	Choice set	Choice	Range and Skew
	Vlaev & Seymour (2009)	Choice set and response scale	Price rating	Range
	Walasek & Stewart (2015)	Choice set	Choice	Range
Neurophysiological	Holroyd et al. (2004)	Outcome set	Event-related brain potential	Range
	Kobayashi et al. (2010)	Outcome set	Single-cell recordings	Range
	Mullett & Tunney (2013)	Outcome set	BOLD response	Range
	Rigoli & Rutledge et al. (2016)	Choice set	BOLD response	Range
	Rigoli & Friston et al. (2016)	Choice set	BOLD response	Range
	Tremblay & Schultz (1999)	Outcome set	Single-cell recordings	Range

Yeung & Sanfey (2004)	Outcome set	Event-related brain potential	Range

Background distribution manipulations. Whereas Ungemach et al. (2011) retrospectively checked participants' receipts to assign participants to the inside-range or the outside-range context, other studies specifically manipulated the background context to investigate how people evaluate alternatives in comparison to current attribute values from memory. Janiszewski and Lichtenstein (1999) asked participants to indicate what price they expected to pay for a product and how attractive a specific price seemed, after they were exposed to a list of 10 different brands with different prices for a specific product. The prices ranged from $0.74–$1.49 (low-range), $0.74–$1.74 (moderate-range), or $0.99–$1.74 (high-range). Participants judged the attractiveness of a specific market price as least attractive when they were previously exposed to low-range prices, more attractive when they were previously exposed to moderate-range prices, and most attractive when they were previously exposed to high-range prices. Expected price estimates were affected the same way: Participants expected highest prices when experienced prices were high-range prices, and expected lowest prices when experienced prices were low-range prices. Adaval and Monroe (2002) obtained analogous results when they asked participant to evaluate three models of a product. The products were either high in price ($184.75–$197.85), low in price ($127.65–134.75) or all presented models cost $159.65. A target product was rated to be less expensive in the context of high-priced products than in the context of low-priced products. This was true, even if people remembered that the price was higher in the high-priced products condition than the low-priced products condition. In addition, the prices encountered for this product even influenced the judgment of a different product two days later. Both these studies show that an option is evaluated against its alternatives. If the price of an option ranks

low within its context, it is perceived to be more attractive than if it ranks high within its context.

Choice set manipulations. Unsurprisingly, research investigating context effects in valuation of risky prospects and risky choice was conducted well before DbS was formulated. Mellers, Ordonez, and Birnbaum (1992) investigated context and response mode effects by asking participants to rate the attractiveness and indicate buying prices (how much money somebody would pay to be able to play the gamble) of a set of simple binary prospects, whose expected values were either positively skewed or negatively skewed. The authors knew people reacted to response mode (indicating higher buying prices for relatively risky Gamble A than the relatively safe Gamble B, but choosing Gamble B over Gamble A) in a different way (Lichtenstein & Slovic, 1971), but did not know how the effects interacted with the distribution of attribute values. Mellers et al. found higher attractiveness ratings for a gamble in the positively skewed distribution than for the corresponding gamble in the negatively skewed distribution. This is consistent with DbS (and RFT as well), because a given gamble will have a higher rank when many others have a smaller EV (in the positive-skew condition) compared to when few others have a smaller EV (in the negative-skew condition). However, distribution basically had no influence on buying prices with simple gambles, but influenced buying prices in mixed gambles in the same way the attractiveness of prospects was influenced.

Ert and Erev (2013) were interested in loss aversion and manipulated the gains (in ₪=Shequels) in mixed gambles to range either from ₪4 to ₪16 (low-value range) or from ₪10 to ₪22 (high-value range). They found that a risky option with a 50% of winning ₪10 or losing ₪10 was chosen more often when participants encountered gains

that were lower in the mixed gamble (in the low-value context) than when ₪10 was the lowest encountered gain (high-value context). Other across trial context effects are reported recently in Rigoli, Rutledge, Dayan, and Dolan (2016). In a new choice task, using four blocks of trials, the authors asked participants to choose between a sure option and a risky option with a 50% zero outcome and 50% double the sure outcome. In two blocks, the risky option comprised gains between £1-£5 or £0, in the remaining two blocks, the risky option comprised gains between £2-£6 or £0. Prior to each block, participants were informed about the range of values. They found that some individuals risked more with increasing amounts, and some risked more with decreasing amounts. Those who risked more with increasing amounts also gambled more with choices that were larger within the context. And those who risked more with decreasing amounts in turn, gambled more with choices that were lower within the context. This means that in both the Ert and Erev (2013) and the Rigoli et al. (2016) studies, participants behaved as if the subjective value of an equivalent choice was larger in the low-value context (₪4-₪16-context and £1-£5-context) than in the high-value context (₪10-₪22-context and £2-£6-context). The DbS explanation would be that a target amount has a higher rank position among many smaller and few higher amounts, than among many higher and no or few smaller amounts.

These findings are also consistent with Vlaev, Seymour, Dolan, and Chater's (2009) findings when estimating the value of pain: Participants paid more to avoid electric shocks in the condition where they experienced a lower-pain sequence as compared to a higher-pain sequence. Additionally, participants paid double the price to avoid identical pain intensity when the range of prices from which a sale price was to be

drawn was doubled. Shifting the range of stimuli to be evaluated, or manipulating the response scale (see more response scale manipulations below) will shift the rank position of target stimuli and influence the willingness to pay, respectively. If somebody experiences the strongest electric shock in a sequence and is asked to indicate how much money she would pay to avoid it, it makes sense that she is willing to pay the highest price. The highest price is going to be higher when the range is doubled, which will lead to a willingness to pay a higher price in this condition.

Just by manipulating the range of gains and losses, Walasek and Stewart (2015) were able to reverse loss aversion, which is often assumed to be an individual difference with a corresponding neural representation (Rick, 2011). In a choice task where participants could always choose to accept or reject a 50% chance of gaining or losing an amount, they found that participants were more likely to accept an offer with an EV of zero if the range of losses was equal to the range of gains than if the range of losses was smaller than the range of gains. And more surprising, participants were more likely to accept offers with negative EVs if the range of losses was double the range of gains than if the range of losses was half the range of gains, reversing loss aversion. DbS explains this phenomenon based on the rank positions of attribute values: Given a choice to accept or reject the offer of a 50% chance of winning or losing $10, a loss of $10 if other losses range from $5–$40 has a lower rank than winning $10 if other gains range from $5–$20. According to DbS this should lead participants to accept this offer. Conversely, people should reject the same offer if the losses range from $5–$20 and gains range from $5–40, because a loss of $10 ranks higher than a gain of $10. If the amounts are not evaluated according to their absolute values, but according to their rank positions, a choice receives

a positive subjective EV if a negative low rank and a positive high rank are offered. As Walasek and Stewart (2015) show, this can lead to phenomena like loss aversion, but also to the opposite.

Vlaev, Chater, and Stewart (2007) were interested in how valid these context effects were in a real world context and chose to ask participants to make choices concerning retirement income provisions. They manipulated the range and the skew of attribute values of saving or investment choices. People chose less risky savings and investments when offered mainly low-risk alternatives with only one high risk alternative, and chose mainly high-risk alternatives when offered mostly high risk alternatives and only one low risk alternative.

Response scale manipulations. In his experiment Birnbaum (1992) tested whether asking participants to make choices instead of judgments would reduce context effects. He presented participants with a gamble. Underneath the gamble there was a list of sure amounts of money (either positively skewed or negatively skewed) in ascending order. Participants were asked to circle all sure amounts that they preferred over the gamble. Consistent with the above findings, more people selected the sure amounts to the gambles when the offered alternative amounts were positively skewed than when they were negatively skewed. If presented with a gamble of a 5% chance to win $48 and otherwise $0 and as an alternative a list of 25 sure gains with more amounts below $20 than above$20, then $20 is looking more attractive than when it is presented within a list of few amounts below $20 and many amounts above $20. Thus, the gambles were worth less in the positive-skew condition than in the negative-skew condition, which is reflected

by people selecting a lower sure amount within a positively skewed context than within a negatively skewed context.

Similarly, in Stewart, Chater, Stott, and Reimers (2003) participants rated gambles also by indicating certainty equivalents (CE) with one set CEs with low values and the other set with high values. What the authors found was that in general, participants chose the CEs that were in the middle and not at the high or lower end of the range, respectively. Note that the effects reported here as well as in Birnbaum (1992) are within trial effects and not across trial effects as when the set of choices is manipulated. Mazar, Koszegi, and Ariely (2013) elicited reservation prices using an incentive compatible procedure, where in some experiments participants were presented with a distribution of prices and in other experiments participants learned the price distribution over time through repeated choices. The price distributions were either positively or negatively skewed. When the price distributions were negatively skewed, participants were willing to pay more money for a certain good than if the price distributions were positively skewed. The DbS explanation would be that the rank position of a certain amount, for example $20, is higher within a price set of prices with a positive skew than within a price set with a negatively skew. Therefore $20 seems like more money in the positively skewed price set than in the negatively skewed price set, making participants willing to pay more money when prices are negatively skewed than when they are positively skewed. There is a similar experiment where reservation prices were elicited and participants experienced the price distribution (with either a low or high range) through drawings out of a bingo cage with similar results (Bohm, Linden, & Sonnegard, 1997).

In a choice task (with a gain or loss domain or across domains), where participants were asked to choose between a fixed gamble and a list of safe choices, where the smallest and the largest safe options were fixed but the amounts of the intermediate safe options varied between conditions, participants were more risk averse when the intermediate amounts were close to the lower end of the range, i.e., positively skewed than when the intermediate amounts were close to the higher end of the range, i.e., negatively skewed (Beauchamp, Benjamin, Chabris, and Laibson (2012).

Benartzi and Thaler (2001) offer a real world example for similar context effects. They obtained data from a propriety database with retirement saving plans and different investment options. There was always at least one safe option and one risky option available, so investors could always select an option to match their true preference–if they had one. They found that as the proportion of risky options increased, the proportion of allocations to risky options increased also. In another real world scenario, Haggag and Paci (2014) found that when taxi drives varied default tip suggestions on their payment terminals to include relatively high values as opposed to relatively low values, people tipped more. All these results are consistent with the rank hypothesis: When participants encountered a positively skewed context in which they evaluated a prospect by choosing a certainty equivalent, indicating a buying or reservation price, or giving an attractiveness rating, the mean CE, buying or reservation price is smaller than when participants encountered a negatively skewed context.

Neurophysiological evidence

In addition to behavioral data, there is research using data from brain imaging on humans or single cell recordings from animal brains, investigating value representation,

which too shows context dependent activation in regions associated with value computation. In an experiment where participants were presented blocks of either low (10p, 20p, and 30p) or high amounts (£5, £7, and £10) of money, which they could win by pressing a button as quickly as possible, using fMRI scans, Mullett and Tunney (2013) found activity in the ventral striatum (VST) and the thalamus, sensitive to the relative value of prizes within one block: The neural response to the highest amount in the low-values block was stronger than to the highest amount in the high-values block, despite the £5 having a much higher objective value than the 30p. Another interesting finding here was that whilst the striatum and the thalamus responses were context dependent within blocks, by the time participants experienced both blocks of the high and the low values, the ventromedial prefrontal cortex (vmPFC) and the anterior cingulate cortex responded with activation proportional to the rank of all experienced stimuli. Although the stimuli are far from uniformly distributed (10p, 20p, 30p, £5, £7, and £10), activation followed ranks of values, not absolute value. In addition to the behavioral evidence above, Rigoli et al. (2016) also collected fMRI data to examine the relationship between the behavioral effects and neural activation. In a within-subjects block design, they asked participants to choose between a sure option and a risky option. In the risky option, the chance of winning double the amount of the sure option was 50%, otherwise £0. In two blocks, the range of gains in the risky option was between £1-£5 or £0 (low-value context), or £2-£6 or £0 (high-value context) in the remaining two blocks. They found that some participants did adapt to the different ranges, others did not. The ventral tegmental area/substantia nigra (VTA/SN) activation differences were conditional upon whether people were context sensitive or insensitive. Context-sensitive participants did not have a

higher activated VTA/SN between the high-value and low-value context, reflecting context adaptation and rescaling of subjective value. However, context-insensitive participants did have a higher activation in VTA/SN in the high-value context than the low value context. In a different study Rigoli, Friston, and Dolan (2016) additionally found that higher activation in the hippocampus at the start of a block is related to more context-sensitivity in the VTA/SN. Although expected value was equal for the safe and risky options, activation in the vmPFC indicated differences between the chosen and the unchosen option and at the same time VST activation pointed to non-linear mapping between the absolute amounts of rewards and their subjective values (Rigoli et al., 2016). Holroyd, Larsen, and Cohen (2004) and also Yeung and Sanfey (2004) asked participant to choose between two cards and learn which was associated with monetary gains and which was associated with losses. They manipulated the range of gains and losses and found that the error-related negativity (an event-related brain potential) depended on the position of the current outcome within the range of other possible outcomes. All these findings suggest internal rescaling and a relative representation of available attributes or outcomes in the brain.

There is more evidence also from animal studies, where the range of outcomes was manipulated to either be wide or narrow. Kobayashi, Pinto de Carvalho, and Schultz (2010) planted electrodes into two monkeys' orbitofrontal cortices and showed that neural firing adapted to different distributions of outcomes, so that the sensitivity to the different prizes changes dependent on the distribution. Using single cell recording in the macaque orbifrontal cortex, Tremblay and Schultz (1999), using separate blocks with food which the macaques liked and food that they did not like, found higher firing rates

for an outcome that they liked in the light of an outcome that they disliked, compared to when the outcome they liked was paired with an outcome that they absolutely loved (see also Elliott, Agnew, & Deakin, 2008 and Seymour & McClure, 2008 for a comprehensive review on relative coding of value in the brain).

We have presented over 20 studies that manipulated the range of possible outcomes or attribute values (most behavioral studies and all the neurophysiological studies were of this type) and some studies that manipulated the skew of the distribution of outcomes and attribute values. The manipulation was either across or within trials and investigated its impact on judgment and choice. The results show that valuation is context-dependent and this sensitivity in valuation is correlated with neural firing and activation of reward related regions in the brain. Changing the range or the skew of possible outcomes can lead to shifts in the rank position of target outcomes within the current sample. The relativity of outcomes or attribute values can be reconciled with a DbS account, where attributes are evaluated according to their rank position within other available attributes from memory. A rank-based evaluation of attributes and alternatives can lead to a reversal of loss aversion, to high prices being perceived as more attractive than comparably lower prices, a willingness to pay more money to avoid the same intensity of pain, higher tipping in taxis, or even more risky choices regarding money allocations in retirement saving funds.

Stewart & Reimers (2008)

There are several yet unpublished experiments that Stewart and Reimers (2008) ran prior to the SRH (2015) experiments, investigating choice-set context effects, which yield results that are in line with the rank hypothesis. In the first series of experiments,

participants rated gambles, where the authors either varied the skew of the probabilities (SR1A) or the distribution of amounts (SR1B). They found that gambles were rated as more attractive if participants experienced a positively skewed distribution of either probabilities or amounts than if participants experienced a negatively skewed distribution of either probabilities or amounts. In the second series of experiments (SR2A-SR2E), the range of probabilities or amount or both were varied and (under certain conditions) participants reversed their preferences on the critical common question, depending on the distribution of outcomes and/or probabilities from previously answered questions. Note that just like in SRH (2015) the attributes and not the expected values were varied.

Experiments SR1A and SR1B

Method. There were 21 participants in SR1A and 32 participants in SR1B. Both experiments were not incentivized and participation was for course credit at the University of Warwick using an undergraduate subjects' pool. Participants were presented with a simple gamble of the type "p chance of winning x" and rated its attractiveness on a scale ranging from 1 ("not attractive at all") to 7 ("very attractive").

SR1A manipulated the distribution of probabilities 1%, 2%, 5%, 10%, 20%, and 80% in the positive-skew condition and 20%, 80%, 90%, 95%, 98%, and 99% in the negative-skew condition. The probabilities were crossed with the amounts £200, £400, £600, £800, and £1,000 to create 30 different gambles for each condition. The two critical probabilities 20% and 80% occurred in both conditions and occupy the highest two ranks in the positive-skew condition whereas they represent the lowest two ranks in the negative-skew condition. We tested whether gambles entailing 20% or 80% in the positive-skew condition were ranked higher than in the negative-skew condition.

According to the rank hypothesis, this should be the case.

SR1B had the exact same design, apart from manipulating amounts instead of probabilities. The five probabilities 5%, 20%, 50%, 80%, and 95% were constant across conditions and amounts £10, £20, £50, £100, £200, and £800 for the positive-skew condition and £200, £800, £900, £950, £980, and £990 for the negative-skew condition were manipulated between participants. Here, we tested whether gambles with a £200 or a £800 amount were rated higher in the positive-skew condition than in the negative-skew condition.

Participants were told that a particular gamble was to be understood as an urn containing 100 balls out of which one ball is drawn. They could click on a radio button to indicate how attractive they found the presented gamble on the current trial using a 7-point scale. After they clicked on Next, the following gamble appeared on the screen.

Results. We compared mean attractiveness ratings for gambles with probabilities 20% and 80% across the two conditions in SR1A and gambles with amounts £200 and £800 across the two conditions in SR1B. Figure 4.1 shows that the distribution participants experienced influenced the attractiveness ratings: Attractiveness ratings for all critical attributes were higher in the positive-skew conditions compared to the negative-skew conditions, $MD_{20\%}$ = 1.83 95% 95% CI [1.41, 2.32], $MD_{80\%}$ = 1.64 95% CI [1.12, 2.16], $MD_{£200}$ = 1.19 95% CI [0.57, 1.81], $MD_{£800}$ = 1.25 95% CI [0.61, 1.88]. The results support the rank hypothesis.

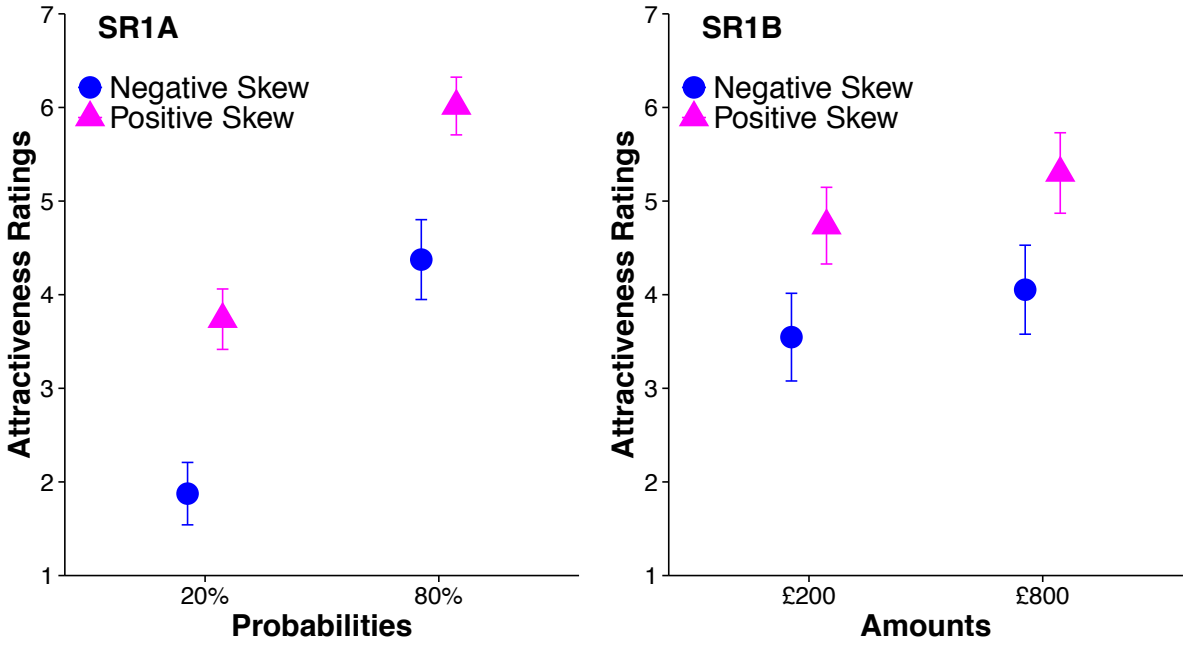

Figure 4.1. Mean attractiveness ratings with 95% confidence intervals for experiments SR1A and SR1B. Gambles containing the critical attributes in the positive-skew conditions were rated consistently higher than in the negative-skew conditions.

Experiments SR2A to SR2E

Method. In five experiments conducted online via Maximiles, Stewart and Reimers (2008) have presented participants with 20 (Experiment SR2A) or 10 trials (Experiments SRB to SR2E) with two gambles to choose from on each trial. The gambles were of the sort of (A) a 75% chance of £25 or (B) a 25% chance of £75. In each experiment, the authors compared the probabilities of choosing gamble A or B on the common questions. As in our experiments, the context (induced by different distributions) was set with the trials leading up to the second to last (critical) trial.

As above, participants were told that a particular gamble was to be understood as an urn containing 100 balls. Two buttons with the probabilities and outcomes displayed on them represented the different gambles. The left and right buttons were set up to

contain the relatively risky versus safe gamble in a random order. Participants made a choice by clicking on the button after which the next gamble appeared. In contrast to the SR1A and SR1B, participants played for real money. Half of the trials were picked at random and played out. Winning 100ipoints was equivalent to winning £1.

SR2A. In this experiment the authors presented participants with 20 questions with the common question being A: a 75% chance of 25 ipoints or B: a 25% chance of 75 ipoints (*N*=98). In Probabilities-Together-Amounts-Apart 75% and 25% were made to look similar by adding 5%, 15%, 85% and 95%. 75% being the third highest and 25% the fourth highest probability. In the same condition 25 ipoints, 35 ipoints, 45 ipoints, 55 ipoints, 65 ipoints, and 75 ipoints were used to make 25 ipoints and 75 ipoints look different in rank (lowest versus highest). According to the rank hypothesis, people should prefer gamble B, because combined with similarly ranked probabilities only amounts should matter. Because 25 ipoints is very different to 75 ipoints in rank, gamble A should appear inferior. In Probabilities-Apart-Amounts-Together, however, probabilities were 25%, 35%, 45%, 55%, 65%, and 75%, making 25% more different in terms of their rank position than 75%. The presented amounts were 5 ipoints, 15 ipoints, 25 ipoints, 75 ipoints, 85 ipoints, and 95 ipoints, making 25 ipoints and 75 ipoints appear very similar to each other. In this condition participants should have a preference for the relatively safe gamble 75% of £25.

SR2B. The design in SR2B was identical to SR2A apart from presenting participants with different distributions of amounts and probabilities (*N*=100). The goal was to avoid suspected strong preferences for one gamble independently of the experiment. All amounts and probabilities are presented in Table 4.2. The common

question was between gamble A: a 30% chance of 100 ipoints and gamble B: a 40% chance of 75 ipoints. Analogous to SR2A, Probabilities-Together-Amounts-Apart should induce more choices in favour of gamble A, the relatively risky gamble. Probabilities-Apart-Amounts-Together should produce more choices in favour of gamble A, the relatively safe gamble.

SR2C. The experimental set-up was exactly the same apart from holding the probability distribution constant with varying amount distributions between conditions (*N*=399). The critical common choice was between gamble A 30% chance of 100 ipoints and B: a 40% chance of 75 ipoints. Amounts-Apart should boost choices for gamble A. Amounts-Together should boost choices for gamble B.

SR2D. SR2D is identical in design and common question to SR2C and SR2B (*N*=200). This time all amounts were kept constant. The probability distributions were varied so that Probability-Together should make gamble A, the risky gamble more appealing. Probabilities-Apart should make the relatively safe gamble B more appealing.

SR2E. The design was identical to SR2C except that gambles were losses (*N*=174). The common trial was choosing between A: 30% chance of losing 100 ipoints and B: a 40% chance of losing 75 ipoints. The hypothesis follows the same logic, but because the experiment entails losses, Amounts-Together should boost gamble A, the relatively risky gamble. Amounts-Apart, on the other hand, should boost the relatively risky gamble B.

Results. Because we had data of multiple experiments we decided to run a random-effects meta-analysis using the metafor package in R (Viechtbauer, 2010) and to estimate the effect size of the context effect. We chose to include all experiments ran by

Stewart and Reimers, although they varied in terms of the domain and the manipulated attributes. This allowed us to obtain the most conservative estimate of the distribution effect size in choice using this type of experimental set-up.

We used odds ratios of the probability of choosing safe divided by the probability of choosing risky. We left out any possible moderating variables, because we did not have enough experiments to use. Figure 4.2 shows the odds ratios for each of the five studies conducted and their 95% confidence intervals. The overall OR is 1.59, 95% CI [1.22, 2.06], which indicates that participants were 1.59 times more likely to choose the relatively safe option, when they are in the safe favoring condition as opposed to the risk favoring condition.

Table 4.2

Experimental set-up in all five SR2 experiments.

Experiment	**Trials**	**Manipulation**	**Amounts**	**Probabilities in %**
SR2A (Gains)	20	Amounts and Probabilities	25,35,45,55,65,75 5,15,25,75,85,95	5,15,25,75,85,95 25,35,45,55,65,75
SR2B (Gains)	10	Amounts and Probabilities	75,80,85,90,95,100 25,50,75,100,125, 150	10,20,30,40,50,60 30,32,34,36,38,40
SR2C (Gains)	10	Amounts only	75,80,85,90,95,100 25,50,75,100,125, 150	10,20,30,40,50,60 10,20,30,40,50,60
SR2D (Gains)	10	Probabilities only	25,50,75,100,125, 150 25,50,75,100,125, 150	10,20,30,40,50,60 30,32,34,36,38,40
SR2E (Losses)	10	Amounts only	75,80,85,90,95,100 25,50,75,100,125, 150	10,20,30,40,50,60 10,20,30,40,50,60

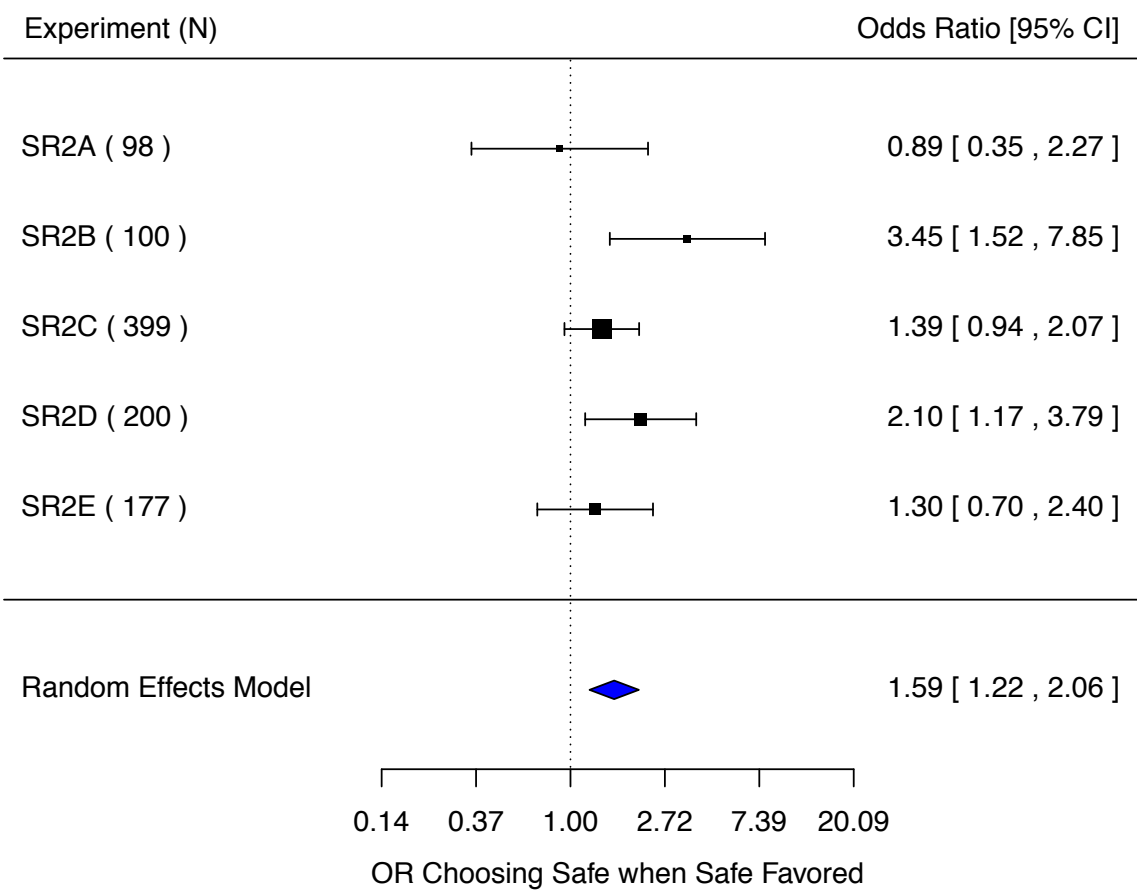

Figure 4.2. Odds ratios with 95% confidence intervals for experiments SR2A to SR2E and the estimated overall effect size (blue diamond with width representing the 95% Cis). Experiment features described in Table 2.2.

The meta-analysis suggests that people's preferences can be reversed by changing the distribution of attribute values. When the presented amounts appeared close to each other in rank, participants tended to choose the safe option more often. When amounts appeared different in rank, on the other hand, more participants chose the risky option. This result also supports the rank hypothesis.

Conclusion

Although in Chapter 2, we cast doubt on the true origin of the context effect observed (when participants learn about distributions by sequentially choosing between

two gambles), the experiments presented here show that the observed context effects are not just revealed through differences in the utility functions (which we now know are partially caused through using the wrong model to fit the data), but also show through differences in attractiveness judgments and choice proportions.

Just as in Chapter 2 or Stewart, Reimers, and Harris (2015), by manipulating the skew of probabilities and amounts in experiments SR1A and SR1B, Stewart and Reimers (2008) demonstrate that attractiveness ratings are lower for gambles with attribute values that are negatively skewed than for attribute values that are positively skewed. Correspondingly, in experiments SR2A to SR2E choices were more risk averse when attribute values were close in rank, whereas they were more risk seeking when attribute values were different in rank. Thus, given these results we can still conclude that valuation and choice are, or at least appear to be rank dependent.

5 Affective evaluation of monetary outcomes is unaffected by rank position

Kassam, Morewedge, Gilbert, and Wilson (2011) asked participants to evaluate outcomes in the light of forgone outcomes and presented an intriguing result: Participants are only sensitive to the absolute amount they win when the prize is the lower of two possible prizes, not the higher. This suggests an asymmetry in the way people evaluate gains. For us, that is particularly interesting, because Kassam et al. hypothesized that this asymmetry depends on working memory and strongly implicates a sampling-and-comparison process much like the one proposed in decision by sampling (Stewart, Chater, & Brown, 2006). In this chapter, we present a replication of Kassam et al.'s basic results, and a reanalysis of their original data. We then present further experiments that test a possible decision-by-sampling-like account of the mechanism underlying their effect. We find a clear cut null result–a null effect of the distribution of attribute values–which we think is an interesting boundary condition for the effects reviewed in the Chapter 4.

Kassam, Morewedge, Gilbert, and Wilson's (2011) experiments

Kassam et al. (2011) presented participants with a card upon which were two latex panels. Each panel covered a sum of money. The participant was asked to scratch one of the two panels, whichever they liked, to reveal their prize. Then they were asked to scratch off the remaining panel to reveal the alternative now forgone prize–the prize they would have won had they only made a different initial choice of which panel to scratch off. After they scratched off both latex panels, participants were asked to report how happy, regretful and disappointed they were about the amount that they had won, on a

scale from one (labeled *not at all*) to seven (labeled *extremely*). These three responses were combined into one positive affect index.

Each participant was handed one of three possible types of cards: Cards had either $1 and $3 prizes, $3 and $5 prizes, or $5 and $7 prizes. The authors hypothesized that if people won the higher of the two prizes (winners) their positive affect ratings would be independent of the absolute prize value. They suggested that the reason for this is that people stop the comparison process after a satisfactory comparison. People would therefore stop after the first and most salient comparison (to the alternative prize, which is lower), which will be satisfactory. For example, if people won $5 but could have won $3, if they had only scratched the other panel first, people make an initial comparison between $5 and $3. This comparison is favorable, so they no longer make further comparisons. From this follows an insensitivity to the absolute prize value for winners. Whichever card people are given, if they scratch to reveal the higher amount, the higher amount will win one comparison (to the lower, forgone amount). If affective evaluation is based on comparisons as Kassam et al. argue, then, because there is no difference in the comparisons, there will be no difference in the affective valuations.

Kassam et al. had a different hypothesis for people who scratch off the lower of the two prizes first. These participants were hypothesized to be sensitive to the absolute prize value. This is because people's most salient comparison (again the alternative prize, which is higher) would not be satisfactory. This unsatisfactory comparison will lead participants to invest cognitive resources to try and find a more satisfactory comparison value. For example, if people won $5 but could have won $7, if they had only scratched the other panel first, their first most salient comparison between the $5 prize and the

forgone $7 prize will make them unhappy. They would therefore make comparisons with other values that come to mind. It is this stage that makes participants sensitive to the absolute value of their prize. The higher their prize, the more likely it is that it will be compared favorably against the other values that come to mind. Comparing their prize to any stable set of values from outside the experiment, for example, should generate positive affect ratings that increase with the absolute value of the prize.

Figure 5.1 plots this interaction effect from Kassam et al.'s Experiment 1. Following Kassam et al. we label conditions with the prize won first and then the forgone prize in brackets. So "$1 ($3)" means a prize of $1 was revealed first, followed by a second and thus forgone prize of $3. Winning the higher of the two prizes (winners) resulted in the same and high positive affect ratings for $3, $5 and $7 prizes (left panel). Winning the lower of the two prizes (losers) resulted in lower positive affect ratings for $1, than for $3 than for $5 prizes. In addition, and unsurprisingly, participants reported an overall higher positive affect when they won the higher of the two amounts compared to when they won the lower of the two amounts.

Kassam et al.'s Experiment 2 hints at their proposed mechanism of why and how this interaction effect emerges. Following their hypothesis, it should be cognitively costly to engage in further comparison processes after the initial, most salient comparison with the foregone prize is unsatisfactory. This is because, when participants lose, they must recall or form some further expectation (or distribution of expectations) against which to evaluate their low prize. To test this hypothesis, Kassam et al. conducted a study where they invited participants in the laboratory and asked them, over multiple trials, to complete a computerized version of the scratch card game. Each participant saw some

filler trials and four critical trials. During the critical trials, participants either had to remember a two-digit number (Low Load) or an eight-digit number (High Load). These trials were rigged so that no matter what box the participants clicked on, they would end up getting the lower of the two amounts (either $3 instead of $5 or $5 instead of $7). If participants were really engaged in an effortful comparison process when they received the lower of the two amounts, imposing a low cognitive load should not impair the comparison process. Thus, under low load participants should remain sensitive to the absolute value of their prize. On the other hand, when participants were under a high cognitive load, they should not be able to engage in the additional comparison process. Thus, in the high load condition, the sensitivity to the absolute value of the prize should disappear.

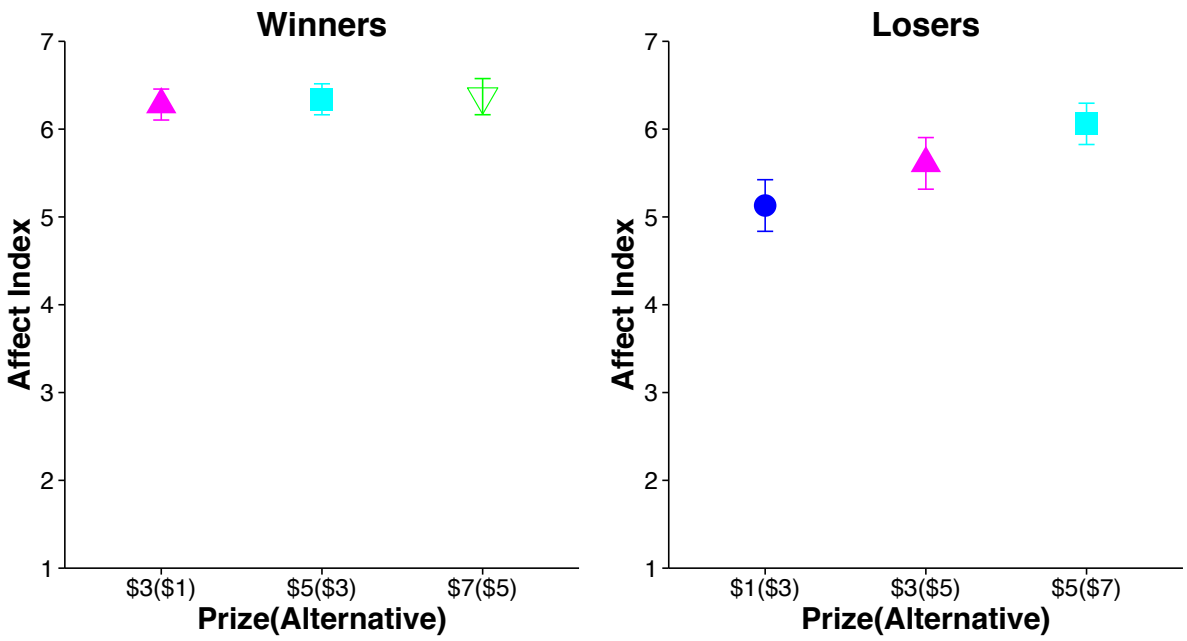

Figure 5.1. Redrawn mean ratings to different prizes with 95% confidence intervals from Kassam et al.'s Experiment 1.

Kassam et al. found a Load-by-Prize interaction with participants being sensitive

to the absolute value of the prize in the low load condition but not the high load condition (Figure 5.2). This experiment replicates Experiment 1's sensitivity to prize within subjects in the low load condition. In the high load condition, in contrast, participants' affect ratings were independent of the absolute prize value. Kassam et al. concluded (a) that people's affective valuations of outcomes do not depend only on the received outcomes themselves, but also on whether the (salient) alternative was higher or lower, i.e. on people's relative status and (b) that the cognitive load manipulation might point to effortful comparisons as one important process involved in the affective valuation of outcomes.

This last result is intriguing; Whilst in Chapter 3 we have unsuccessfully attempted to show working memory capacity as a key factor in the decision making process, Kassam et al. appear to have shown that there has to be a link: Finding that the sensitivity to the absolute size of the prize disappears with high cognitive load, implies working memory is an important part of the mechanism for valuation–at least when participants are engaged in the further evaluation of their prize, prompted by receiving the smaller of the two prizes. This finding is what led to us further exploring working memory load in the experiments below. We return to this issue when we introduce our experiments. First, however, we review the broader literature for differences in processing of good outcomes (e.g., winning the higher of two prizes) and bad outcomes (e.g., winning the lower of two prizes).

Differential processing of positive and negative outcomes

In the early days of the judgment and decision making literature, Galanter and Pliner (1974) reported that a decrement results in a disutility exponent that is larger than

the utility exponent for an equal increment. Following that, a core idea in prospect theory is loss aversion–that losses loom larger than gains–which manifests itself a steeper value function in the loss domain (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991).

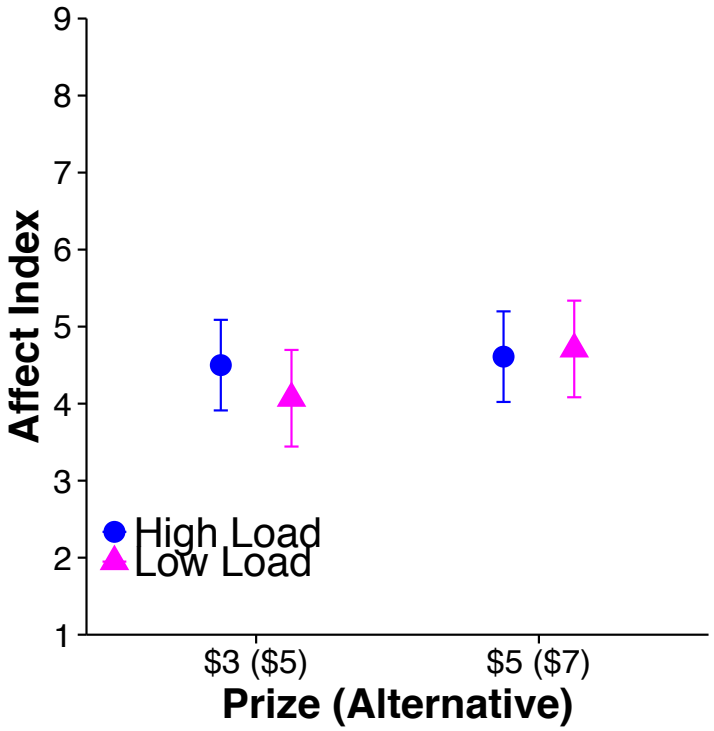

Figure 5.2. Redrawn mean ratings to prizes with 95% confidence intervals in critical trials from Kassam et al.'s Experiment 2. We have presented the y-axis to cover the full 1-9 range. It demonstrates the effect of cognitive load on the sensitivity to the absolute amount of the prize.

There are many demonstrations of negativity having a greater impact on valuation, especially regarding attitude, social interaction and impression formation (Baumeister, Bratslasky, Finkenauer, & Vohs, 2001; Lewick, Czapinsi, & Peeters, 1992; Peeters & Czapinski, 1990). Rozin and Royzman (2001) identified many additional domains where negativity bias manifests itself and offered several theories–generally based on the adaptation principle–to account for the vast occurrence of negativity bias

and the positive-negative asymmetry.

In a recent paper, Sakaguchi and Stewart (2016) analyzed tweets with negative and positive connotations, and found that negative tweets were more likely to include words that imply causal thinking. Causal thinking requires engagement with a negative event. This finding is in line with research in social cognition showing that people reason the most when they experience negative events. One reason why this could be the case is that reasoning could serve as mechanism to control negative feelings (Abele, 1985). Another way to regulate negative feelings and behavior could be to make satisfying comparisons (Epstude & Roese, 2008). Pyszczynski, Greenberg, and LaPrelle (1985) for example observed that participants who were led to believe that they performed poorly on a bogus test (a negative event) were more likely look for information about the performance of other participants when they expected other participants to have performed poorly. Kassam et al. (2011) hypothesized that dealing with negative outcomes is more likely to lead to an effortful search for satisfying comparisons than it is when dealing with positive outcomes and they presented data in line with this hypothesis.

What Kassam et al.'s studies also show is that equally high prizes can be perceived as positive or negative events in the light of a foregone alternative. The salient alternative can make a prize seem rather bad, if the alternative is larger, or rather good, if the alternative is smaller, which the authors showed led to different evaluations and possibly also processing of the exact same outcome. The fact that salient alternatives matter, and impact valuation, is also in line with decision affect theory (Mellers, Schwartz, Ho, & Ritov, 1997) and regret theory (Loomes & Sudgen, 1992). In both of

these theories, the psychological value of an alternative is based, in part, on the comparison with the alternatives under other states of the world.

These counterfactual comparisons offer a possible resolution of the Easterlin paradox: at the level of nations, despite increasing income, wellbeing has not increased. Clark, Frijters, Shields (2008) used social comparison and habituation to explain the Easterlin paradox: An increase of everybody's income in the same time frame, does not lead to an increase in individual happiness, if people compared their income with the income of their peers. Increases in your own income do not make you happy when everyone else's income increases at the same time (see also Boyce, Brown, & Moore, 2010).

The role of salient comparisons and the question about which alternatives matter was also posed by Medvec, Madey and Gilovich (1995) in a study where they investigated emotional responses of Olympic medallists. Their results suggested that bronze medallists tended to look happier than silver medallists did. Their explanation for this effect was that all medallists think of what might have happened, and compare their current state to these could-be scenarios: silver medallists might have won gold, whereas bronze medallists might have not won a medal at all. The different standards of comparisons of silver and bronze medallists would cause bronze medallists to be happier than silver medallists. Research by Larsen, McGraw, Mellers and Cacioppo (2004) supports this explanation: When they asked participants to rate a win that could have been better on a positive and a negative affect scale, they found that participants experienced both, positive and negative emotions simultaneously. This is perhaps due to

making satisfying and dissatisfying comparisons at the same time. The current state might determine which comparisons dominate.

Reanalysis of Kassam et al.'s Experiment 2: Winners too love money

Kassam et al. kindly provided us with the raw data from their Experiment 2. We have exactly replicated their analysis of the responses to the four critical trials, as described above. In addition to replicating the analysis Kassam et al. report in their paper, we also report an additional analysis here; an analysis to the responses to these filler trials. While participants made affective ratings in all critical trials, they only made ratings in two of the seven filler trials. We used those for our analysis. In one trial participants always won $5 instead of $1. In the other trial participants always won $3 instead of $1. Recall that Kassam et al. only compared trials with a gap of $2 between prize and alternative. We however, calculated the difference in positive affect to winning $5 rather than $1 when the alternative prize was $3. Whilst Kassam et al. found that participants were equally happy when winning $5 instead of $3 as they were when winning $3 instead of $1, suggesting that positive affect does not increase with the absolute prize, we found that positive affect does increase when the difference between the prize and the alternative amounts increases, Mean difference (MD) =1.71 95% CI [1.27, 2.16] (see Figure 5.3). Thus given their data, we might slightly modify Kassam et al.'s claim: Winners do not just care about winning–they also care by how much they win.

Experiment 1

Because of the importance of Kassam et al.'s effect for the series of experiments that we planned, we began by collecting new data in an attempt to replicate their Experiment 1. Recall that in this experiment Kassam et al. found that participants'

positive affect was sensitive to absolute amount only when the alternative was higher, but insensitive to absolute amount when the alternative was lower (given that the gap between the prize and alternative was equal).

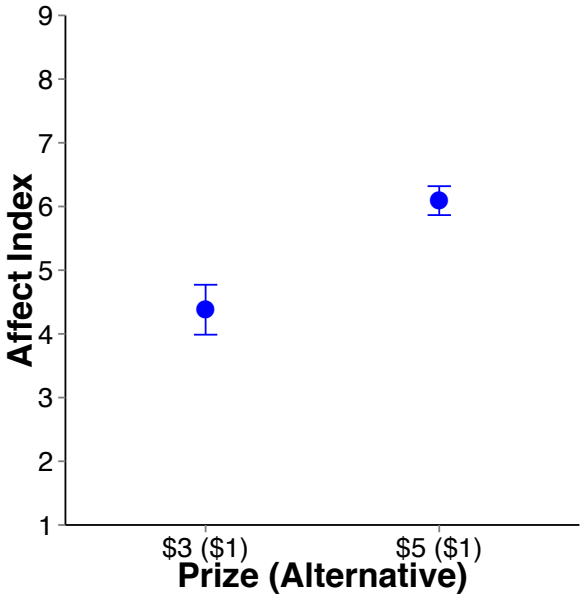

Figure 5.3. Means and 95% confidence intervals of positive affect for two filler trials of Kassam et al.'s Experiment 2. The difference between the prize and its alternative clearly matters for the valuation of the prize.

Method

Participants and Design. 432 participants located in the USA were recruited using Amazon Mechanical Turk. The departmental ethics committee approved this and all subsequent studies and all participants gave informed consent to take part by proceeding from the introduction to the main part of the experiment. To replicate the amount sensitivity experiment, we used a 3($1 and $3, $3 and $5 or $5 and $7)-x-2(winning the lower vs. winning the higher amount)-between-subjects design, in which we a) varied the amounts revealed to the participants when they clicked on the card and b) whether the prize value was lower (loser) or higher (winner) than the alternative amount. Participants

received a participation fee and the prizes were hypothetical sums of money.

Procedure. In a slight variation of Kassam et al.'s original task, we had participants click on one of two cards on a computer screen, rather than scratching one of two latex panels on a single card. Participants went through 1 trial of choosing cards and answering three questions. They were asked to choose one out of two cards, which hid an amount that was revealed by clicking on the card. After the amount was revealed and it said "You win $x", participants were asked to click on the remaining card and find out what the alternative amount was. When they clicked the other card, it said: "You could have won $y". For the winners, the first revealed amount was always higher than the alternative, for losers the first revealed amount was always lower than the alternative. After they have seen what is hidden under both cards, their prize and the foregone alternative, we asked them to indicate on a scale from one (not at all) to seven (extremely) how happy, regretful (reverse coded) and disappointed (reverse coded) they felt after having won the prize money. As in the original experiment we aggregated the three measures to one positive affect index.

Results

Out of 432 participants we deleted 33 participants because of duplicate IP addresses. We calculated means and confidence intervals of the positive affect index to compare the groups within and between factors. We replicated the results of Kassam et al.'s Experiment 1. As Figure 5.4 shows, when participants win the lower of the two prizes (in conditions $1($3), $3($5) and $5($7)), on a scale from one to seven, their positive affect is highest when they win $5, lower when they win $3, $MD_{\$3\text{-}\$5}$=0.34 95%CI [-0.14, 0.83] and lowest when they win $1, $MD_{\$1\text{-}\$3}$=0.66 95% CI [0.17, 1.16] and

$MD_{\$1\text{-}\$5}$=1.00 95% CI [0.51, 1.50].

However, when participants win the higher of the two amounts (in conditions \$3(\$1), \$5(\$3), and \$7(\$5)), they report being equally happy no matter what the absolute prize money was, with overlapping confidence intervals, $MD_{\$7\text{-}\$5}$ = 0.09 95% CI [-0.18, 0.37], $MD_{\$5\text{-}\$3}$=0.02 95% CI [-0.25, 0.31], $MD_{\$7\text{-}\$3}$=0.06 95% CI [-0.21, 0.34]. Compared to Kassam et al.'s Experiment 1, in our study participants reported higher positive affect in the winner conditions (\$3(\$1), \$5(\$3), and \$7(\$5)) and lower positive affect in the loser conditions \$1(\$3), \$3(\$5), and \$5(\$7)). The here relevant relative status by prize money interaction effect however, replicated perfectly.

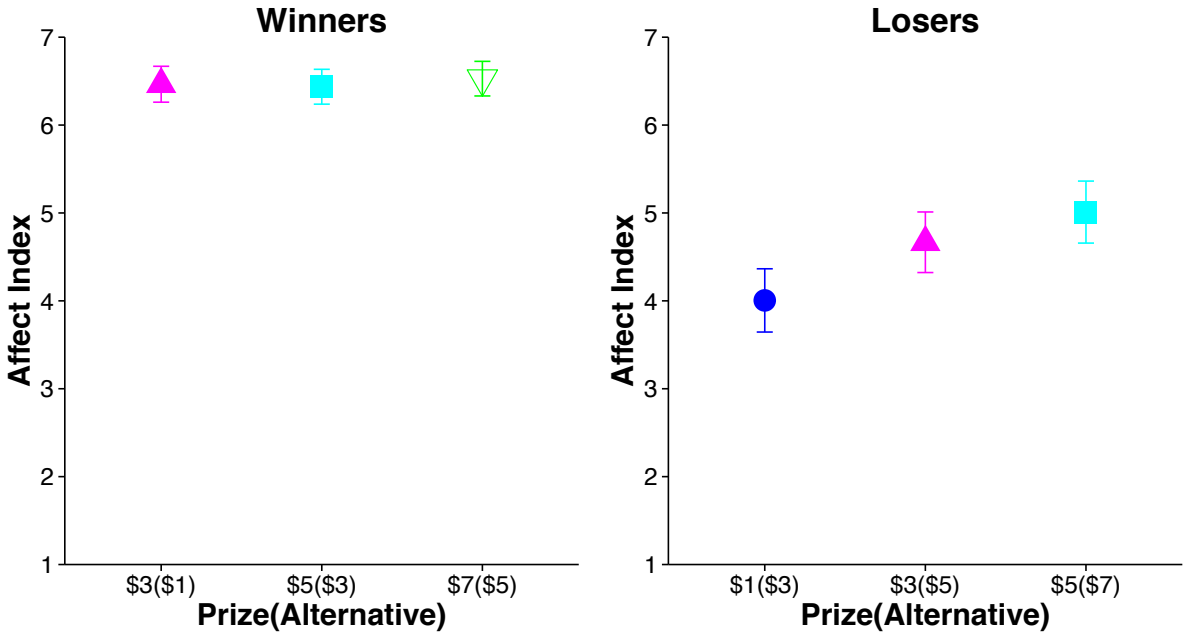

Figure 5.4. Means and 95% confidence intervals of positive affect after learning about the prize money and its alternative separately for winners and losers in Experiment 1. The same color and shape indicates equal prize money.

Testing a decision by sampling account of the Kassam et al. effect

Here we describe a possible decision-by-sampling account of the mechanism by which the losers show sensitivity to the absolute value of their prize. We only focused on losers, because this is where the DbS account fits in well with Kassam et al.'s explanation of their effect. Given that we find that Kassam et al.'s winners also seem to be sensitive to the alternative prize value, this does suggest that a similar hypothesis could be drawn for winners and losers alike.

Decision by sampling assumes that values are constructed during a series of binary, ordinal comparisons between the target value and the other values in working memory. These other values are assumed to be both the values present in the immediate context, such as the forgone alternative value behind the other panel in these experiments here, and other values recalled from long-term memory, such as expectations about other possible prizes or memories of other sums of money, which will be taken into account if they are diagnostic (Gilbert, Giesler, & Morris, 1995). What follows from this binary, ordinal comparison process is that if values are constructed by counting up favorable comparisons, values of attributes will be based on the rank position of the target value within the current sample of values in working memory. That is, the number of favorable comparisons will be proportional to the number of values in the sample smaller than the target value. We call this the rank hypothesis. Because sampling values from long-term memory is cognitively costly (Morewedge, Gilbert, Myrseth, Kassam, & Wilson, 2010), a working memory load would disrupt the comparison process between the current prize and other prizes from memory (but not to the salient forgone prize).

There is evidence to support the rank hypothesis of judgment in many different fields. We know that wage satisfaction is rank-dependent, and that general life satisfaction can be predicted by the rank position of somebody's income within a comparison sample and that job satisfaction depends on the relative rank of one's income among incomes of one's colleagues (Boyce, Brown & Moore, 2010; Brown, Gardner, Oswald & Qian, 2008; Card, Mas, Moretti & Saez, 2011). We also know that judgments of symptom severity of people with depression, pain ratings, and gratitude all depend on and can be predicted by relative rank (Melrose, Brown & Wood, 2013; Watkins, Wood, Lloyd & Brown, 2013; Wood, Brown & Maltby, 2011). Aldrovandi, Wood, and Brown (2013) even showed that judged crime severity and the according punishment is not affected by the actual crime severity itself, but where it is believed to lie within a social comparison.

Could it be that the sensitivity to the absolute value of the prize for losers found by Kassam et al. is not actually revealing the sensitivity to the absolute amount per se, but instead is the result of extending a decision-by-sampling comparison process beyond the forgone prize to a larger sample of values from memory? This wider sample would provide a stable background against which a relative comparison process could operate to produce what looks like sensitivity to absolute value. In this account, the process by which winners rate their happiness higher than losers (and higher the bigger the difference between their prize and the alternative), would be the same process by which losers' happiness ratings are sensitive to the prize. For winners, comparisons are made to the lower forgone amount, but for losers, comparisons could be even more expanded, to all other possible amounts from outside the current pair of outcomes.

The aim of the following experiments is to test this possibility. We manipulated the history of prizes participants experienced in a between-subjects design. Thus, when the participants encountered a critical losing trial, they will have recently encountered different distributions of prizes. If the decision-by-sampling account is right, then we should be able to see differences in the affect ratings on the losing trials across the different distributions of prizes participants have encountered on earlier trials. This would be evidence that it is the relative rank and not the absolute value of an amount that is predictive of positive affect. Specifically, by making people experience several different outcome values before the last critical outcome, we can test whether positive affect, after winning a prize is higher when its rank is high within its sample than if it is low within its sample.

In contrast to our findings in Chapter 3, Kassam et al. provided evidence that working memory load lead to within-participant changes in valuation. This is why we made a second attempt at investigating this factor by adapting Kassam et al.'s design. We followed the procedure of their Experiment 2 and although their effect size for cognitive load was small, we also manipulated working memory load to investigate how it can play into the valuation process. If the decision-by-sampling account is correct in the assumption that working memory is required for the series of binary, ordinal comparisons from which value is constructed, we should see working memory load eliminate or at least reduce the effect of the different distributions of prizes. By inducing cognitive load, attribute values from previous trials would be either pushed out of participants' working memory and/or the comparison process would be impaired. Such an elimination would be evidence that working memory is an important prerequisite for the comparison and

accumulation process.

Experiment 2

The next three experiments all build on the fact that participants' positive affect differs between prize moneys when they win the smaller of two possible prizes and that salient comparisons influence judgment. The next three experiments test two hypotheses at the same time: First, given that people were sensitive to the absolute amount of the smaller prize, are they also sensitive to the distribution of other foregone prizes and amounts encountered throughout the experiment? Specifically, we are interested in whether the rank hypothesis, which tests whether positive affect depends on the rank position of a prize within other prizes on offer holds when evaluating outcomes. To test this, we manipulated the distributions of prizes before the critical trial: In one between-subjects condition, the prize won on the critical trial would be the lowest seen so far and the prize forgone on the critical trial would be the highest seen so far (Fillers-Inside). In this case, we expect the prize won to be perceived as much lower than the forgone prize and thus receive a very low rating. In the other between-subjects condition we manipulated the distribution of prizes such that there were prizes smaller than the prize won on the critical trial and that there were prizes higher than the foregone prize. Now the critical-trial prizes will fall right in the middle of the distribution of prizes from earlier trials, with the fillers either lower or higher (Fillers-Outside). In this case, we expect the prize won to be perceived as quite similar to the forgone prize and thus receive a higher rating than in the other condition (see Figure 5.5).

The second hypothesis we are testing is, as Kassam et al.'s propose, that prize sensitivity arises from the comparisons that people make. If it is true that people are

sensitive to the distribution of prize values and thus winning the prize in Fillers-Inside receives a lower rating than winning the same prize in Fillers-Outside, then, according to Kassam et al.'s hypothesis, sensitivity should be reduced by high cognitive load. As in Kassam et al., we will compare sensitivity to the distribution between low and high cognitive load ratings. DbS assumes that working memory plays an important role for evaluation as it implies and allows the sampling from memory and pairwise comparison process. By manipulating cognitive load, we test the role of working memory in the process leading up to rank effects in judgment. If we found the rank hypothesis confirmed and also found that it is removed or reduced by high cognitive load, this would suggest comparisons with a reference sample from memory (as proposed by DbS).

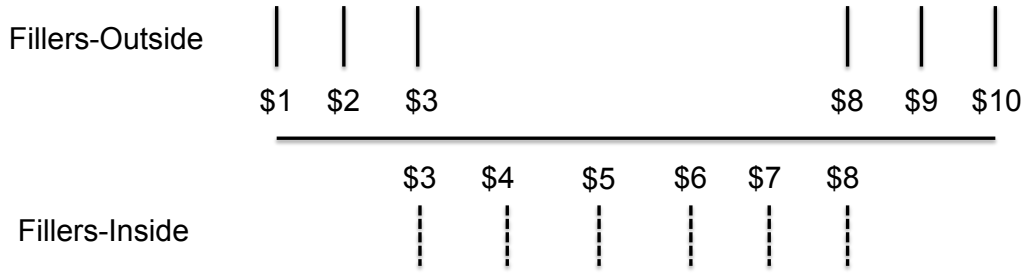

Figure 5.5. Example distribution of prizes in Experiment 2A. $3 and $8 are the prizes on offer on the critical trial.

Experiment 2A

Method. 430 participants took part in this online experiment that we ran via MTurk, same as all other experiments. During 15 trials of the same type as in Experiment 1 we varied the distribution of amounts participants experienced. The critical 15th trial was fixed to be "You win $3" and "You could have won $8." In one manipulation, we chose the filler amounts for the first fourteen trials to make the $3 and $8 seem similar. Specifically, participants experienced the distribution Fillers-Outside with prizes $1, $2,

$3, $8, $9, $10 where $3 and $8 have similar rank positions, rank 3 and 4. In the Fillers-Inside manipulation the distribution entailed prizes $3, $4, $5, 6$, $7, $8, where $3 and $8 have very different rank positions, rank 1 and 6. Crossing all amounts resulted in 15 trials per condition, of which the first 14 served as fillers, making $3 and $8 appear similar (Fillers-Outside) or different (Fillers-Inside). Additionally, we varied cognitive load randomly throughout the first 14 trials. For the last trial, we set cognitive load per condition: before the trial started, either an 8-letter string (High Load) or 2-letter string appeared on the screen and had to be memorized during the valuation task. After that participants were asked to recall the string. This experimental set-up resulted in a 2 (Fillers-Inside vs. Fillers-Outside) x 2 (Low Load vs. High Load on the last trial) design. We were interested whether in the last –the critical–trial, participants reported a lower positive affect in Fillers-Inside than in Fillers-Outside and whether these differences were moderated by cognitive load. To avoid ceiling effects, the scale was adapted to reach from one to nine, as in Kassam et al.'s Experiment 2.

Results. After removing participants with duplicated IP addresses, as we planned in advance, 402 participants remained for the analysis. We first checked whether memorizing eight instead of two letters was indeed more challenging by checking the accuracy rates between conditions. Participants in the low load conditions were better at memorizing the letters than did participants in the high load conditions, MD = .32 95% CI [.24, .41].

As Figure 5.6A shows, participants' positive affect was neither sensitive to cognitive load, nor to the distribution of amounts experienced before the critical trial, MD_{Low}= 0.06 95% CI [-0.52, 0.63] and MD_{High}= -0.15 95% CI [-0.77, 0.46]. We find, at

most, a small effect of the distribution of prizes experienced across both the high load and the low load conditions, $MD_{Overall}$= -0.04 95% CI [-0.46, 0.37], while Kassam et al. and we found a 1-point difference on a 1-7 scale between losers receiving $1 and losers receiving $5. Our best estimates of the effect the distribution of amounts is about 0.1 of a point on a 9-point scale, with CIs indicating it is unlikely that the effect of distribution is larger than 0.5 of a point.

We realized that having all prizes as single digit numbers strongly suggests that they are drawn from a uniform distribution of single digit $1-$9 (or maybe $1-$10) prizes, and this prior could overshadow the more subtle experimental manipulation. In Experiments 2B and 2C, we used different distributions of two digit numbers, sometimes across the whole range, and in some comparisons only small two digit numbers to avoid activating only numbers across the whole range (all two digit numbers). We think this will help avoid participants having a strong prior expectation of a uniform distribution, which may be responsible for the zero or very small effect in Experiment 2A.

Experiment 2B

Method. 624 participants took part in Experiment 2B. The experimental set-up was exactly the same as in Experiment 2A. Here we used three distributions of amounts. Amounts $13, $29, $46, $65, $81, and $96 were used for Fillers-Outside and $46, $49, $53, $58, $61, and $65 for Fillers-Inside. These two distributions share the prizes $46 and $65 and serve as critical test between these conditions. We added the condition Fillers-Inside-Small with amounts $13, $16, $19, $22, $26, and $29 to compare with Fillers-Outside using the “You win $13” and “You could have won $29” as the critical trial. This will serve as an additional test of whether the distribution manipulation worked by using

prizes that are at the lower end in Fillers-Outside.

In contrast to Experiment 2A, the gaps between all prizes were fairly equally spaced within one condition. The critical trials for the comparison between Fillers-Outside and Fillers-Inside were the 15th trials, "You win $46" and "You could have won $65". The critical trials for the second comparison, which also serves as a manipulation check, was the 15th trial in Fillers-Inside-Small, "You win $13" and "You could have won $29" and whichever trial was the one where participants won $13, with forgone $29 in Fillers-Outside. This critical trial is never the last one, but randomly appeared throughout the course of the experiment in Fillers-Outside. Everything else remained the same as in Experiment 2A.

Results. After removing participants with duplicated IP addresses, 587 participants remained for the analysis. We again checked whether participants in the low load condition were more accurate in the memorization task: Similar as above, participants in the low load conditions were better at memorizing the letters than did participants in the high load conditions, MD= .29 95% CI [.22, .36].

As Figure 5.6B shows, participants' positive affect was neither sensitive to cognitive load, nor to the distribution of amounts experienced before the critical trial, MD_{Low}= 0.46 95% CI [-0.14, 1.05] and MD_{High}= 0.24 95% CI [-0.35, 0.84].

The second comparison however, does show an effect of distribution: Under low and high cognitive load participants report a higher positive affect in Fillers-Outside than in Fillers-Inside-Small, MD_{Low}=0.70 95% CI [0.01, 1.39], MD_{High}=1.28 95% CI [0.61, 1.95] (See Figure 5.6B–$13($29)). Although high cognitive load even seems to increase distribution sensitivity, there is no interaction effect, $MD_{Load*Distribution}$=0.58 95% CI [-

2.98, 4.14]. This is the first time across our experiments in this chapter where we find an effect of the distribution of prizes from earlier trials.

To test whether this effect was real, we needed to replicate it. We also wanted to consider cognitive load. Maybe the effect of distribution so far was attenuated by inducing cognitive load on all trials leading up to the critical trial. It is possible that because participants experienced some amount of cognitive load on every trial, the different distributions could not be properly encoded. Thus, we might expect that choosing prizes without needing to memorize letters on earlier trials might give a stronger effect of those prizes during the last, critical trial.

Experiment 2C

Method. 421 participants took part in the last experiment. We copied the distribution of the fillers-outside and the fillers-inside conditions. In this experiment all but the third and the last trial were without cognitive load. The reason why we included load on the third trial was so participants already knew what it would look like and were prepared for the procedure when they arrived at the critical trial. In comparison to Experiment 2B, we additionally dropped the third condition pair and so had two distributions crossed with two load conditions just as in Experiment 2A. Everything else remained the same.

Results. We removed participants with duplicated IP addresses, so that 400 participants remained for the analysis. As before, but with a higher accuracy overall– probably due to only including two instead of all 15 trials with cognitive load– participants in the low load conditions were more accurate in the memorization task than participants in the high load conditions were, *MD*= .23 95% CI [.16, 30].

As Figure 5.6C shows, and in line with all but one of our experiments, participants' positive affects were very similar across conditions (MD_{Low}= 0.06 95% CI [-0.53, 0.65] and MD_{High}=0.20 95% CI [-0.35, 0.76]). In contrast to Experiment 2B–$13($29), the rank of the alternative prize did not impact people's positive affect ratings.

To summarize so far, the ratings in the low-load-outside conditions were hypothesized to be higher than those in the low-load-inside conditions, because the alternative prize in the fillers-inside conditions created a larger difference in rank position between the won and forgone prizes. The effect of distribution was only found in 2B–$13($29), but did not replicate. The difference was hypothesized to be attenuated by cognitive load, but given we see no effect of the distribution of earlier prizes, we cannot expect it to be attenuated by cognitive load and indeed we did not see this.

Meta-analysis

We have conducted three experiments, with four comparisons for both low and high cognitive load conditions. Our pattern is fairly consistent across experiments with the exception in Experiment 2B–$13($29). To be able to estimate the overall effect size for the manipulation of the distributions from our four comparisons and to compare the effects from across our four comparisons with one another (and with other experiments), we conducted a meta-analysis using mean differences between the fillers-inside and the fillers-outside conditions. Because the experiments were not exact replications of one another we implemented a random-effects model and used cognitive load as a moderator variable.

As Figure 5.7 shows, the effect of the distribution manipulation overall is a difference of 0.32 95% CI [-0.02, 0.65] on a 9-point scale. That is, the difference in the

rank positions of the prizes has a positive effect of perhaps one third of a point on a nine-point scale. Given the confidence interval, the effect is probably positive, but it could be as small as zero or it could be as large as two thirds of a point.

Our manipulation of rank was not subtle: In the fillers-inside condition the difference between the prize and the alternative was the difference between receiving the lowest (1st) and highest (6th) ranking outcomes in the experiment; in the fillers-outside condition, the difference was between two adjacent ranks (3rd and 4th) in the middle of the distribution. For comparison, in our reanalysis of Kassam et al.'s Experiment 2 fillers, the difference in ratings between \$5(\$1) and \$3(\$1) was 1.71 points 95% CI [1.27, 2.16] on the same nine-point scale. The difference between \$1(\$3) and \$5(\$7) for the losers in our Experiment 1 was 1.00 point 95% CI [0.51, 1.50] on a seven-point scale. The effect of the distribution of prizes that we have found in our four comparisons is small in comparison to the Kassam et al. effect sizes.

Given we at best estimated a small effect of the distribution of previous prizes, we should not expect cognitive load–which we hypothesized should reduce this difference between the distribution of attribute values–to have a large moderating effect. And it does not: Cognitive load has no or only a very small effect on the difference in ratings on the critical trial across conditions, with wide confidence intervals on the estimate, $beta_{Load}$ = 0.07 95% CI [-0.65, 0.78].

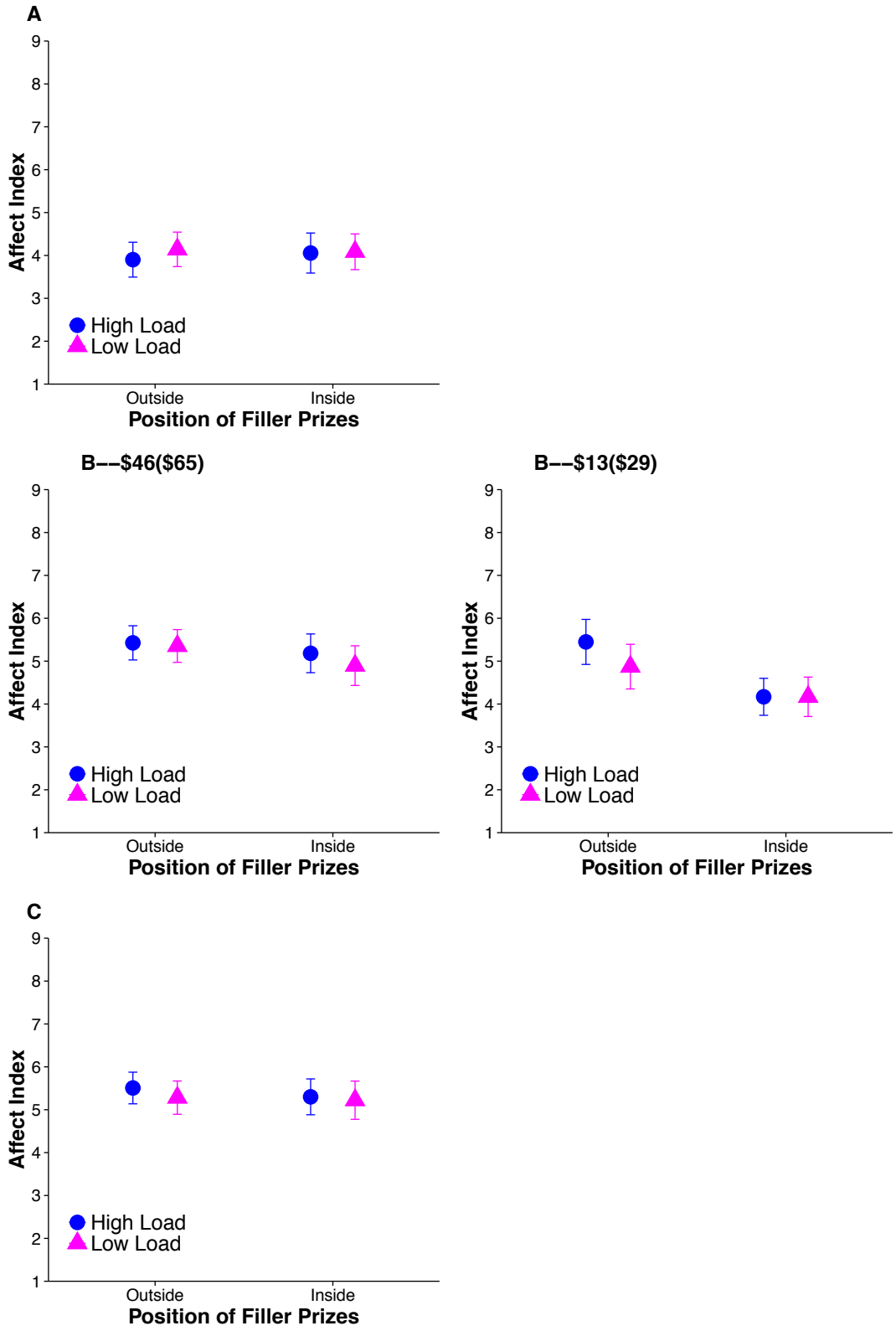

Figure 5.6. Means with 95% confidence intervals of positive affect ratings for the critical

trials in Experiments 2A-2C. Generally, there is no effect of either distribution or cognitive load across experiments.

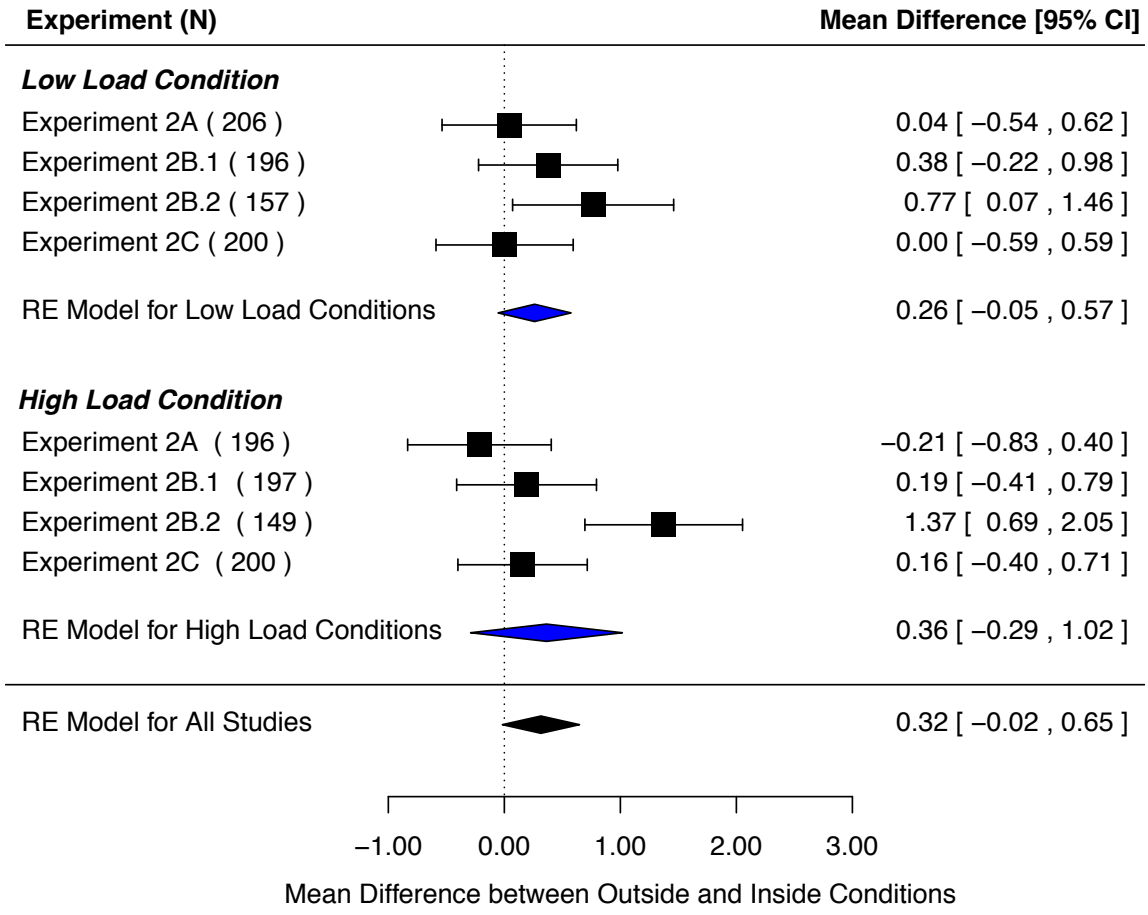

Figure 5.7. Mean differences with 95% confidence intervals between Fillers-Outside and Fillers-Inside, separately for low and high cognitive load of Experiments 2A-2C.

Conclusion

We have replicated Kassam et al.'s finding that people are sensitive to the absolute value of a prize only when it is the smaller of two possible prizes but not the larger. We also manipulated the rank position of the prize and the forgone alternative by varying the prizes experienced in earlier choices, so that the selected and forgone prizes would be very different or very similar in rank between participants and hypothesized

that the bigger the difference in rank the lower people's reported positive affect would be. Additionally, inspired by Kassam et al.'s findings on the effect of cognitive load on people's amount sensitivity, the second goal was to test whether the hypothesized rank effect would decrease under high cognitive load, which would be in line with DbS. However, because we failed to find an effect of rank, we cannot estimate how cognitive load reduces this effect.

To contrast these findings, results of very similarly set-up experiments by Stewart and Reimers (2008) yielded comparably big differences on attractiveness ratings of gambles dependent on the rank position of the manipulated attribute, as reported in the previous chapter. One difference between the experiments is that in the Stewart and Reimers experiments, participants evaluated options, whereas here they evaluated outcomes. Other studies however, do find effects when judging their life satisfaction or their wage satisfaction (Boyce, Brown, & Moore, 2010; Card et al., 2012), which in itself are outcomes. Maybe outcomes experienced here are just not aversive and relevant enough for participants to engage with, as opposed to studies where more serious questions are asked. Gilbert, Lieberman, Morewedge, and Wilson (2004) showed that a certain amount of negative affect must be reached before people engage in rationalization. This problem could apply in our case: If participants did not feel bad enough, they will not take the other prizes into account and thus affect ratings will not depend on other foregone prizes. We think our results could point to a boundary condition in terms of when DbS-like processes might lead to rank effects in judgment and when they do not.

6 A Negative Zero is Better than a Positive Zero: The Mutable-Zero Effect and Category-Consistent Counterfactuals

Apparently innocuous changes in the description, or "framing," of options can dramatically reverse preferences between options. One way in which framing can reach its effect is by evoking alternatives or counterfactuals to which options or their features are compared. For instance, in the evaluation of a job offer, a given income can be seen as better if it is framed in order to bring to mind other smaller incomes, like salaries of other junior staff members, than if it brings to mind other higher incomes, like salaries of other senior staff members. Kahneman and Miller (1986) referred to the ease with which alternatives or counterfactuals are brought to mind as the *mutability* of the information given. The *mutable-zero effect* (Scholten, Read, Canic, & Stewart, 2016) is one example of such a framing effect: Swapping the word "receive" zero for "pay" zero increases an option's attractiveness. For example, when choosing between

A: "Receive £100 today and pay £400 in a year" and

B: "Pay £200 today and pay £0 in a year,"

most people chose Option B. However, when choosing between

A': "Receive £100 today and pay £400 in a year" and

B': "Pay £200 today and receive £0 in a year,"

most people chose Option A'. The manipulation is the simplest possible, in that the difference is just one word, but it has a very reliable effect on preference.

Scholten et al. (2016) offered a hypothesis for the effect based on Kahneman and Miller's concept of relative mutability. The terms "pay zero" and "receive zero", they proposed, bring to mind different counterfactuals. The thought of paying something

brings to mind other amounts that one might pay (and zero is the best payment possible), while the thought of receiving something brings to mind other amounts you might receive (and zero is the worst possible receipt). These counterfactual values are important in valuations that arise from the expectations that certain categories and situations form (Kahneman & Miller, 1986). In this paper we develop this idea further, and test three versions of the counterfactuals hypothesis. We call these the *Any-Counterfactuals* Hypothesis, the *Category-Consistent-Counterfactuals* Hypothesis and the *Unhappiness-Induced-Counterfactuals* Hypothesis. These hypotheses all assume that the comparisons that people make with their spontaneously generated counterfactuals are what give rise to the mutable-zero effect. They differ in terms of the conditions under which and what kind of counterfactuals are used for the comparison process. We next discuss the hypotheses in turn.

The Any-Counterfactuals Hypothesis

In the Any-Counterfactuals Hypothesis, the phrases "pay zero" and "receive zero" prompt people to generate different counterfactuals. Once generated, these counterfactuals are treated like any other counterfactual. Just as in the Ebbinghaus illusion, the counterfactuals would work by contrast: Surrounding a central circle with many smaller circles makes the central circle appear larger and surrounding the same central circle with many larger circles makes the central circle appear smaller. In this analogy, the "pay/receive zero" is the central circle, and the counterfactuals are the surrounding circles. So in this hypothesis the mutable-zero effect emerges just because of the comparisons with any available counterfactuals. According to the Any-Counterfactuals Hypothesis, people are influenced not only by spontaneously evoked,

implicit counterfactuals–it is assumed that the phrases “pay zero” and “receive zero” induce different counterfactuals to which zero is then compared–but also by counterfactuals encountered in the environment, which could be explicit counterfactuals added by the experimenter. If the explicit counterfactuals work just as the implicit counterfactuals, then these explicitly presented counterfactuals will work alongside and blend in with the implicit counterfactuals people generate, or even suppress or reverse the effect of those implicit counterfactuals. That is, we will see two additive affects: the first is the effect of the implicit counterfactuals generated by participants upon reading the words “pay zero” or “receive zero”, and the second is the effect of the explicit counterfactuals in the environment. For example, if participants read “receive zero” they will generate implicit counterfactuals about other larger receipts. If they then learn from the environment that “receive zero” could have been any other receipt, these explicit larger receipts will combine with the implicit counterfactual larger receipts to make “receive zero” seem especially bad. But if they instead learn from the environment that the “receive zero” could have been any other payment, these explicit counterfactual payments will combine with the implicit larger receipts, and because the implicit and explicit counterfactuals are acting in opposite directions on the “receive zero”, the effect of the implicit counterfactuals will be attenuated or even reversed. Similarly, if participants learn that “pay zero” could have been any other payment, they will be more likely to choose this option than when participants learn that it could have been any other receipt.

If the data supports the Any-Counterfactual Hypothesis, it would mean that the mutable-zero effect works by comparing and contrasting implicit and explicitly presented

counterfactuals alike. It would also mean that we could manipulate the mutable-zero effect by contrasting the mutable zero with either favorable (all payments) or unfavorable (all receipts) counterfactuals. We will contrast this below with the idea in the Category-Consistent-Counterfactuals Hypothesis that the words "pay" and "receive" in front of zero will constrain the possible counterfactuals that will be used for the comparison process.

The Category-Consistent-Counterfactuals Hypothesis

Kahneman and Miller (1986) entertain "the compelling intuition that some alternatives are closer to reality than others and some changes of reality are smaller than others," and that "a counterfactual possibility should be 'close' if it can be reached by altering some mutable features of reality" (p. 7). Suppose that, when considering "receive zero", only counterfactual receipts encountered in the environment are close to reality, and that counterfactual payments are not, and that when considering "pay zero", only counterfactual payments encountered in the environment are close to reality, and that counterfactual receipts are not. Then, counterfactuals encountered in the environment will only have an effect (or at least will have a larger effect) when they match the spontaneously evoked counterfactuals. Following this framework, only counterfactuals matching the pay/receive category of the mutable zero would be close enough to reality and therefore be valid comparison values. Thus "pay zero" would only be compared to explicitly provided payment counterfactuals, and explicitly provided receipt counterfactuals will have no effect. Similarly, "receive zero" would only be compared to explicitly provided receipt counterfactuals, and explicitly provided payment counterfactuals will have no effect. This Category-Consistent-Counterfactuals Hypothesis

differs from the Any-Counterfactuals Hypothesis in that participants are only affected by counterfactuals in the environment when they match the pay/receive category of the mutable zero, and, unlike the Any-Counterfactuals Hypothesis, participants will be unaffected by counterfactuals where the pay/receive category does not match the mutable zero, because they are not close enough to reality.

The Unhappiness-Induced-Counterfactuals Hypothesis

Negative evaluations demand greater cognitive processing (Kanouse & Hanson, 1987; Peeters & Czapinski, 1990). Thus, in the Unhappiness-Induced-Counterfactuals Hypothesis, it is possible that only "receive zero", but not "pay zero", will ever be compared to counterfactuals encountered in the environment. In Chapter 5 Experiment 1 we replicated Kassam et al.'s experiment implying that, when an initial outcome is satisfying enough, people avoid further sampling and avoid making further comparisons, because they already are in a desirable emotional state. For "pay zero", the initial counterfactuals generated are other payments, and these all make "paying zero" look good—so the comparison process stops because people are already in an emotionally desirable state. Counterfactuals in the environment will not be compared with "pay zero" because the comparison process has stopped. For "receive zero", the initial counterfactuals generated are other receipts, and these make "receiving zero" look bad–so the comparison process continues in an attempt to reach an emotionally desirable state. Now counterfactuals in the environment will be involved in continuing comparisons with "receive zero" and will thus affect the valuation of "receive zero". The Unhappiness-Induced-Counterfactuals Hypothesis thus predicts that counterfactuals provided by the

experimenter will only have an effect (or at least will have a stronger effect) for "receive zero" options, but not for "pay zero" options.

Figure 6.1 shows the designs used in our test of the three hypotheses under investigation. In all experiments, participants make a choice between two options. Before making the choice, a (seemingly) random device determined a missing value on one of the options (either from a spinning fruit machine drum or the turning over of face-down cards). Figures 6.1A and 6.1B show our "spinner" random device which we used in Experiments 1A-B and 2A-B-C. Figures 6.1C and 6.1D show our "cards" random device which we used in Experiments 2D and 3A-B-C. In all experiments, participants viewed various forgone values on the random device before the final value, which was always either "pay 0" or "receive 0", was filled in. One factor was whether the label of the mutable zero as "pay" or "receive", i.e., "pay zero" or "receive zero". The second factor, orthogonal to the first, was the set of foregone values, i.e., the explicitly provided counterfactuals in the environment (context items, for short).

Experiment 1

Method

Design. The participants' primary task was to make one choice between two options: the Static Option "Receive £100 today and pay £400 in a year" or the M0 Option (the option entailing the mutable zero) "Pay £200 today and [receive/pay] £0 in a year". Whether the £0 in the M0 Option was described with the word "pay" or "receive" was manipulated between participants. Initially the £0 component in the M0 Option was not set but was (apparently) determined by a slot machine spinner, which we fixed to finish on either "pay £0" or "receive £0". The spinner was realistically animated to give the

impression of viewing a cylindrical drum with outcomes printed on it rolling and stopping. The view of the drum was constrained by a window, just like on a slot machine. When the spinner stopped, we changed the color of the spinner to match the rest of the option and blend in, which gave the appearance of the "[pay/receive] £0" embedding itself into that option. Participants clicked on a button to activate the spinner. We manipulated alternative entries on the spinner (context items) which spun past to be either "pay £100", "pay £200", and "pay £300" or "receive £100", "receive £200", and "receive £300" between participants. This resulted in a 2 (receive versus pay category) x 2 (receive £100-£300 context items versus pay £100-£300 context items) between-participants design. We have labeled conditions ReceiveContext-Receive0 (for "receive context, receive zero"), ReceiveContext-Pay0, PayContext-Receive0, and PayContext-Pay0. See Figure 6.1 for the experimental set-up and click here, if you would like a go at the experiment yourself.

Participants and procedure. We collected data from 96 Warwick students who participated for course credit (Experiment 1A). Additionally, we recruited 202 students via Prolific Academic (Experiment 1B), a new online platform similar to Amazon Mechanical Turk, but with an English university student population. The departmental ethics committee approved this and all remaining studies. Participants gave informed consent to take part before they started the experiment. In the online version, they were informed that proceeding from the introduction to the main part of the experiment is equivalent to giving consent. For this one-choice task, we paid the online participants £0.60.

Figure 6.1. Experimental set-up of all nine experiments. In 1A and 1B participants click on „Spin“ and watch the context items (in the grey box) pass by before it stops at either „pay zero“ or „receive zero“, at which point the box blends in and turns green. In 1C and 1D participants click on the blue cards in whichever order they prefer. The last card they click on moves into the space where the „?“ was and turns green. Click here to play the experiment with the spinner or here to play the experiment with the cards set-up

In addition to making the choice between the Static Option and the M0 Option, we asked participants to memorize and recall the four entries on the spinner. To ensure that they saw and would remember the entries, they could click on a button labeled "replay the spin" and see the entries spin past exactly as before as many times as they wished. They were instructed that they would have to recall the entries after they made their choice, to make sure they were paying attention. The choice and memorization were not incentivized.

The results would support the Any-Counterfactuals Hypothesis if the M0 Option was more attractive in the PayContext-Pay0 and PayContext-Receive0 than the ReceiveContext-Receive0 and ReceiveContext-Pay0 conditions. On the other hand, the results would support the Category-Consistent-Counterfactuals Hypothesis if the M0 Option was more attractive in the PayContext-Pay0 and ReceiveContext-Pay0 than the PayContext-Receive0 and ReceiveContext-Receive0 conditions. Finally, the results would support the Unhappiness-Induced-Counterfactuals Hypothesis if the M0 Option was most attractive in PayContext-Pay0 and ReceiveContext-Pay0 conditions and more attractive in the PayContext-Receive0 than in the ReceiveContext-Receive0, which would make the M0 Option seem least attractive.

The Any-Counterfactuals Hypothesis predicts only main effects of our manipulation of "pay" or "receive" zero category and our manipulation of "pay" or "receive" context items. The Category-Consistent-Counterfactuals Hypothesis predicts an interaction, such that our manipulation of "pay" or "receive" contexts will only have an effect or a much stronger effect when the pay/receive category of the context matches the category of the mutable zero. The Unhappiness-Induced-Counterfactuals Hypothesis

predicts a different interaction, such that our manipulation of "pay" or "receive" contexts will only have an effect for "receive zero" and not for "pay zero". Figure 6.2 plots out the predictions of the three different hypotheses.

Results and discussion

We calculated the proportion of participants choosing the M0 Option in Experiment 1A and 1B (Figure 6.3.1A and 6.3.1B). We fitted a two-predictor logistic regression to estimate the odds of choosing the M0 Option as a function of whether the context items were payments or receipts, and whether the M0 Option was a payment or a receipt (Table 6.1).

Table 6.1

Odds Ratios (OR) for choosing the M0 Option with 95% confidence intervals.

Experiment	Context-OR [95% CI]	Mutable Zero Category-OR [95% CI]	Interaction-OR [95% CI]
1A	0.98 [0.35, 2.65]	5.50 [2.13, 15.86]	2.78 [0.39, 21.76]
1B	0.99 [0.52, 1.89]	2.95 [1.58, 5.68]	1.11 [0.31, 4.02]
2A	4.31 [1.36, 16.65]		
2B	2.63 [1.67, 6.21]		
2C	1.86 [0.98, 3.55]		
2D	1.80 [1.03, 3.17]		
3A	1.71 [1.12, 2.63]	2.39 [1.57, 3.67]	1.71 [0.73, 3.99].
3B	0.85 [0.57, 1.26]	1.63 [1.09, 2.44]	1.11 [0.50, 2.45]
3C	1.27 [0.85, 1.43]	1.56 [1.04, 2.33].	0.92 [0.41, 2.05]

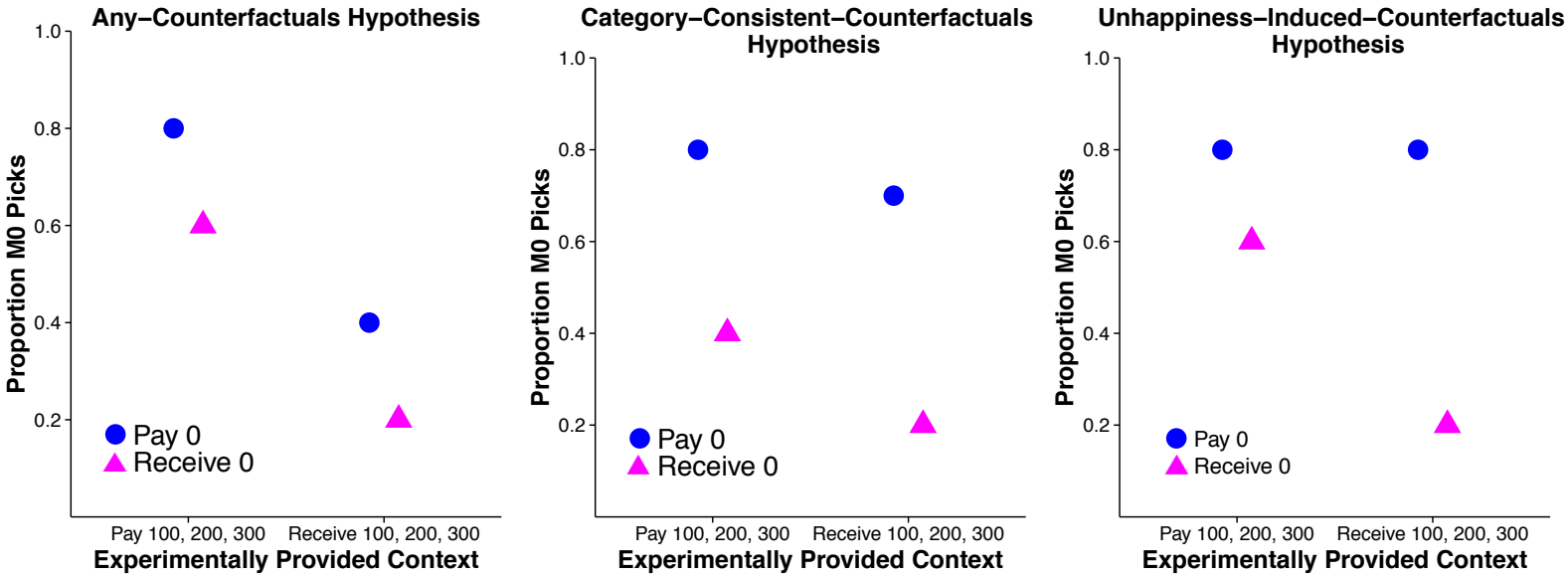

Figure 6.2. Predicted choice proportions of M0 picks predicted according to the three hypotheses.

The main effect of the mutable zero category was the same as in Scholten et al.'s (2016) earlier studies, with more people choosing the M0 Option with "pay zero" like in conditions PayContext-Pay0 and ReceiveContext-Pay0 than with "receive zero" like in conditions PayContext-Receive0 and ReceiveContext-Receive0. The estimate of the main effect of context, given by the values that participants encountered on the spinner, was almost exactly null (i.e., odds ratio about 1), though the relatively wide confidence intervals leave open the possibility of an effect as large as halving or doubling the odds of choosing the M0 Option. Finally, the estimate of the interaction between the factors was poor—the wide confidence intervals in both experiments allow us to say little about whether the effect of the context depended on the label of the mutable zero.

However, another possibility is that people did not attend to the experimentally provided counterfactuals on the spinner. To check if this is so, we ran the next series of experiments. In Experiment 2, we tested whether participants were sensitive to the context when the zero was not labeled with "pay" or "receive" and thus could not obviously belong to either category. The M0 Option was simply presented as "£0" without the words "pay" or "receive" preceding it. We anticipated that if participants paid attention to the counterfactuals encountered in the environment, zero should take on that same category as the counterfactuals from the environment. If it does take on the same category, participants should be less likely to choose the M0 Option in the ReceiveContext condition because "£0" here is relatively unattractive. On the other hand, participants should be more likely to choose the M0 Option in the PayContext condition because "£0" here is relatively attractive. And independently from all of the theorizing about counterfactuals, if the manipulation of context had no effect for unlabeled zeros,

we see little reason to expect an effect for labeled zeros. Thus Experiments 2A-D serve as a check of the potential strength of our context manipulation which, to preempt the results, show robust effects of context.

In Experiment 1 we have not taken the performance of the recall task into account. Because only the mutable zero category and not the counterfactuals encountered in the environment affected participants' choices, for Experiment 2 we decided in advance to look only at those participants' choices who correctly remembered which category the counterfactuals in the environment belonged to. We ran four variants of Experiment 2, as detailed below.

Experiment 2

Method

Design. In Experiment 2 the counterfactuals encountered in the environment (context items) were either all payments or all receipts, and zero was always unlabeled. We therefore had two conditions: the PayContext condition ("pay £100", "pay £200", "pay £300", before settling on "£0") and the ReceiveContext condition ("receive £100", "receive £200", "receive £300", before settling on "£0").

We made an additional change to the recall phase. We gave participants the option between "receive" and "pay" from a drop-down menu, rather than requiring them to type the word before the number. The aim was to make it as easy as possible to recall the label correctly and as intuitive as possible for participants to understand what their task was. Also, the memorization process should make the context items more salient and therefore make the effect–should it exist–as large as possible.

Participants and procedure. We collected data from 124 Prolific Academic

participants for Experiment 2A. We found that only about half of participants could correctly choose the options from the drop-down menu so we ran Experiment 2B as a replication of Experiment 2A. In order not to lose power by excluding too much data, Experiment 2B had 210 Prolific Academic participants, roughly double the number of Experiment 2A. To ensure that participants paid attention to the context items, we added an instruction stressing that it was essential for them to remember the four entries on the spinner. However, we still found poor recall of the context items. We therefore ran Experiment 2C with 186 Prolific Academic participants. In Experiment 2C, in order to be able to make a choice, participants had to fill in the context items correctly. By making sure that participants couldn't proceed until they have correctly filled in the context items, we traded-off being able to use the data of all 186 participants with the possible impact that swapping the order of the recall and the choice task might have had on participants' choice.

For Experiment 2D we ran 201 Prolific Academic participants and used a similar experimental set-up but now, instead of clicking a button to spin a spinner, we presented participants with four cards with hidden amounts that were revealed one-by-one as they clicked on the cards to turn them over (see Figure 6.1C-D). Participants were told that the last card they clicked would fill in the missing space in the M0 Option. Participants clicked on the cards in any order they wanted. Unbeknownst to the participants, the last card they clicked always offered "£0". At the same time the "£0" was revealed, it slid into the missing space and blended into the M0 Option. The motivation behind switching from spinners to cards was to allow the context to remain on the screen at the time of the choice, removing the need for memorizing the category of the context and thus the

cognitive load that might interfere with the choice during the task. At the same time, it increased the salience of the context. Second, participants didn't have to think about a rationale for memorizing the context, which made it seem more intuitive than the spinner version with memorization. We also made a slight change to the amounts on the option, equalizing the sum of outcomes of the two options, which therefore implied a zero interest rate. Here participants chose between the Static Option: "Receive £400 today and pay £900 in a year" and the M0 Option: "Pay £500 today and £0 in a year".

Results and discussion

As we consider remembering if the context was "receive" or "pay" as an easy task, especially because all context items belonged to the same category, we had decided in advance to exclude all participants who did not correctly remember the context category. In Experiment 2A, 48 out of 124 participants did not choose the correct option (either "pay" or "receive") all three times in the memory test. We excluded 23 participants in the PayContext condition and 25 participants in the ReceiveContext condition, which left us with 76 participants. We excluded a similar proportion of participants in Experiment 2B (104 out of 210 participants, 58 from the PayContext and 46 from the ReceiveContext condition).

We had expected some people to fail to recall correctly whether the context category was "pay" or "receive", given the poor recall seen in Experiments 1A and B. It is still surprising to us that so many people failed to recall only three entries from the same category when instructed to do so. It was especially puzzling, because in other unrelated experiments run at the same time also using the subject pool provided by Prolific Academic, participants' choices showed remarkable consistency, suggesting

reliable data coming from the platform we used.

The proportion of participants choosing the M0 Option throughout the four experiments is presented in Figure 6.3.2A-6.3.2D and odds ratios of choosing the M0 Option as a function of the context items ("pay £100", "pay £200", "pay £300" or "receive £100", "receive £200", "receive £300") are displayed in Table 1. Throughout all four experiments, the context consistently influenced choice. The M0 Option was more attractive in the PayContext than the ReceiveContext conditions. From this series of experiments, we also concluded that participants did pay attention to the context and that they did make comparisons between zero and the context items, as the context was the only thing that differed between PayContext and ReceiveContext.

In summary, when the M0 Option is unlabeled, we found evidence for the Any-Counterfactuals Hypothesis. There is an alternative interpretation however: Instead of comparing the zero with the context items, the context category could be applied to zero–as if zero inherits its category from the context–and become a mutable zero, which then would lead to the mutable-zero effect. However, we think this category inheritance option is less likely. In other experiments, entirely in the domain of gains, we find that the context has a large effect, and this cannot be due to the target item inheriting its category from the context category, because the target items' categories do not differ in those experiments.

So far, we consistently replicated the mutable-zero effect in the original form (Experiments 1A and B). We also see a large context effect when the zeros are unlabeled (Experiments 2A-D). Evidence for the Unhappiness-Induced-Counterfactuals Hypothesis, however, is less clear because of the large confidence intervals on the interaction effect in

Experiments 1A and B. Thus, we ran one more series of experiments with the cards set-up from Experiment 2D, in which respondents revealed counterfactual cards one at a time, and the cards remained onscreen when a choice was being made. One advantage of the cards set-up is that we were able to retain the data of all participants. The other advantage is that the cognitive load of the task was minimized so that the memorization of the context items did not interfere with the choice that people made.

Experiment 3

Method

Design. This series of experiments is a combination of the designs of Experiments 1A-B and Experiment 2D. The options were the same as in Experiment 2D and we used the cards set-up, which maximized the salience of the context items. As in Experiments 1A-B, the £0 in the M0 Option was described with the word "pay" or "receive" and the context (displayed as cards were turned) was pay £100, £200, £300 or receive £100, £200, £300, both manipulated between participants. In Experiments 3A-B the expected values of both options were net losses. For Experiment 3C, we decided to test the effects in the gain domain too. So in Experiment 3C the Static Option was "Receive £900 today and pay £400 in a year" and the M0 Option was: "Receive £500 today and [receive/pay] zero in a year". This change resulted from swapping all receive-attributes for pay-attributes and vice versa from Experiments 3A and B. Everything else remained the same.

Participants and procedure. We collected data from 401 participants for Experiment 3A, 398 participants for 3B and 398 participants for Experiment 3C, all via Prolific Academic. As before, we paid participants £0.60. Other details followed the

procedure of Experiment 2D.

Results and discussion

The proportion of participants who chose the M0 Option throughout the three experiments is presented in Figure 6.3A-6.3C and odds ratios of choosing the M0 Option providing evidence for the three hypotheses are displayed in Table 6.1. Participants consistently chose the M0 Option more often when the mutable zero was "pay zero" rather than "receive zero". In Experiment 3A we found that the context had a moderate influence, but the finding did not replicate in the remaining two experiments. Here, the interaction effect was better estimated than Experiment 1A and B, and we can rule out large effects.

The experiments in this last series replicated the mutable-zero effect found in Experiments 1A and B. But we found very little or no evidence for the effect of the counterfactuals encountered in the environment, or an interaction with the label of the mutable zero. Using the cards set-up allowed us to have more power to detect possible effects of the context and the interaction effect also. Should there be such effects, they are probably not very large. Experiment 3C extends the results from dealing with net losses to net gains and does not suggest differences between the two domains.

Meta-analysis

We have run three series of experiments, nine experiments in total. We found consistent effects of the mutable zero category in every experiment where we tested for it (i.e., in Experiments 1A and B, and Experiments 3A-C). However, when the zero was labeled with "pay" or "receive" we find at best weak evidence for the effect of manipulating the context. But in Experiments 2A-D, where the zero was unlabeled, we

find strong evidence for the effect of context items. Figure 6.3 shows the choice proportions pattern for all nine experiments.

To combine evidence over our experiment series, we now use a random effects meta-analysis to estimate the effect sizes for the effect of the mutable zero category, the context items, and their interaction. We used the metafor package in R (Viechtbauer, 2010), with the odds ratios as effect size measures.

Main effect of Mutable Zero

To quantify the effect of the mutable zero, we only included Experiments 1A-B and 3A-C, as Experiment 2A-D omitted the label from the zero. With "pay zero", participants were more likely to choose the M0 Option than with "receive zero", *OR* = 2.19 times, 95% CI [1.41, 3.41] (Figure 6.4). The effect size of the mutable zero category in the zero-labeled experiments (Experiments 1A-B and 3A-C) is comparable to the effect size of context in the zero-unlabeled experiments (Experiments 2A-D).

Main effect of context

To quantify the effect of context, we used all nine experiments. Because, in Experiment 2 the zero was unlabeled, we used Label (with or without) as a moderator. As Figure 6.5 shows, the effect of the context depends on the presence of the label. If there was no label prior to zero (Experiment 2), participants were more likely to choose the M0 Option in the PayContext conditions than the ReceiveContext conditions, *OR* = 2.26 95% CI [1.24, 4.13]. However, the context items did not influence choice when zero already belonged to either the "pay" or the "receive" category, *OR*=1.26 95% CI [0.81, 1.97].

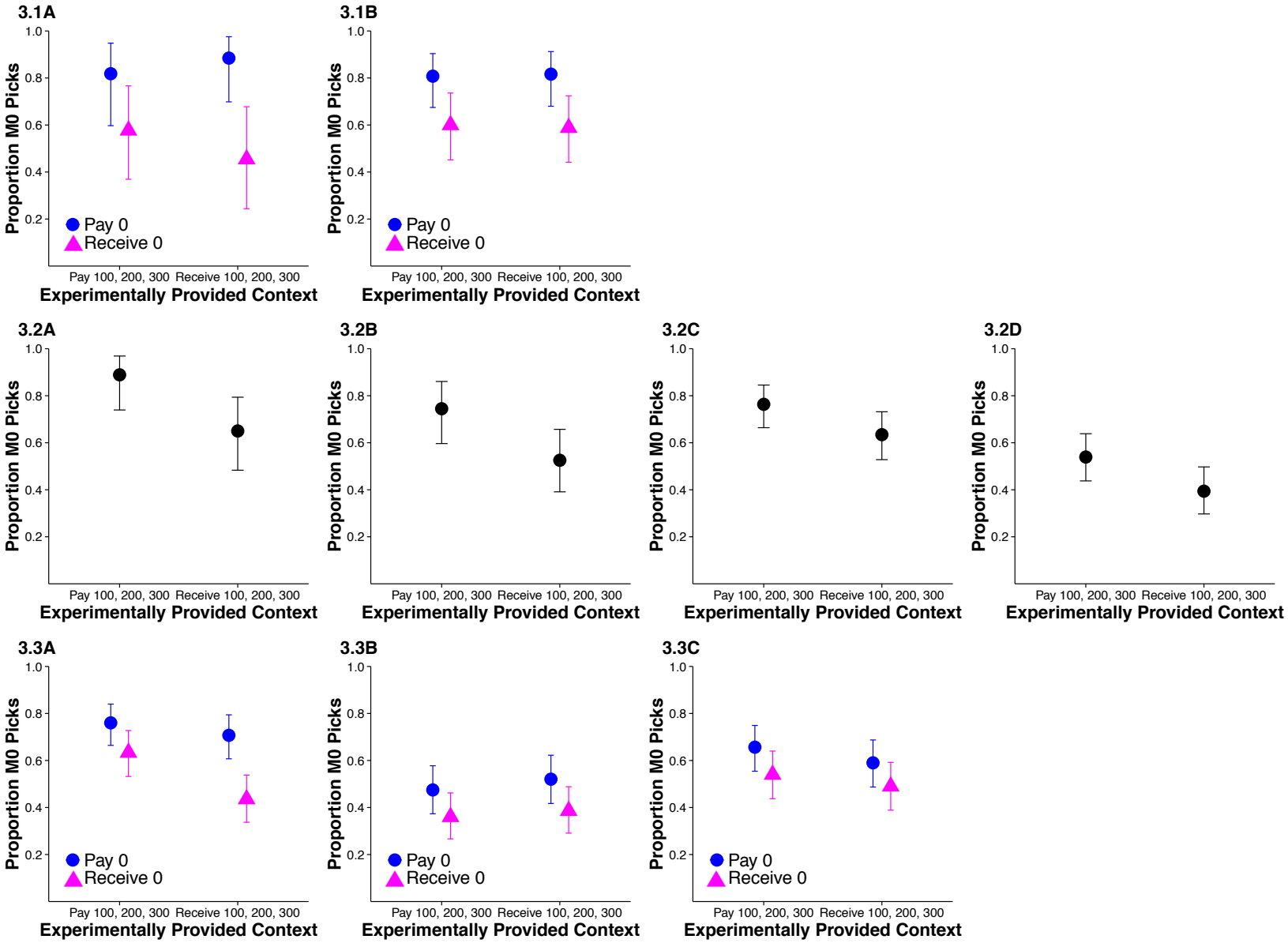

Figure 6.3. Choice proportions of M0 picks and 95% confidence intervals for all nine experiments.

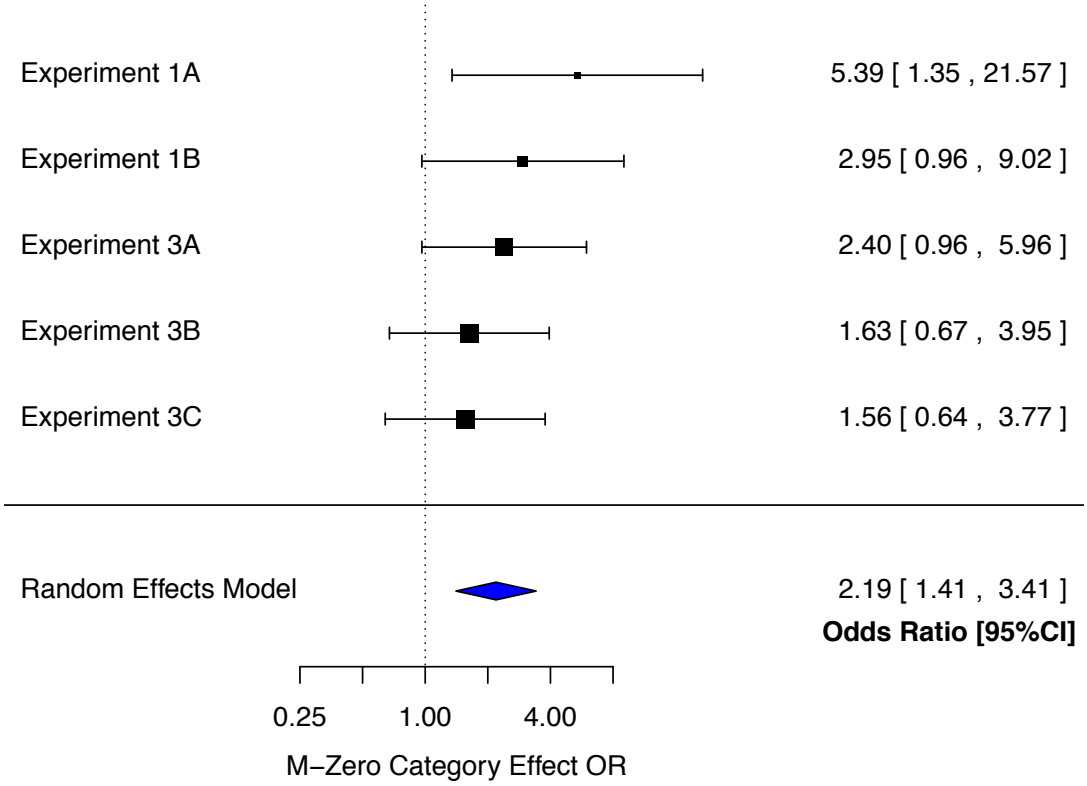

Figure 6.4. Random effects meta-analysis to test the effect of the mutable zero category. Odds ratios (OR) greater than 1 indicate an increase in the preference for the M0 Option with „pay zero“ instead of „receive zero“.

Context-by-M0 interaction

To establish how big the interaction effect is between the final zero category and the context items, we again only used Experiments 1 and 3. Although Experiments 1A and 3A indicated a potential difference between the effect of context when the mutable zero was labeled as “receive zero” from when it was labeled as “pay zero”, when combining all experiments, the meta-analysis yielded only a weak interaction between context and mutable zero category with wide confidence intervals, *OR*=1.29 95% CI [0.69, 2.43] (Figure 6.6).

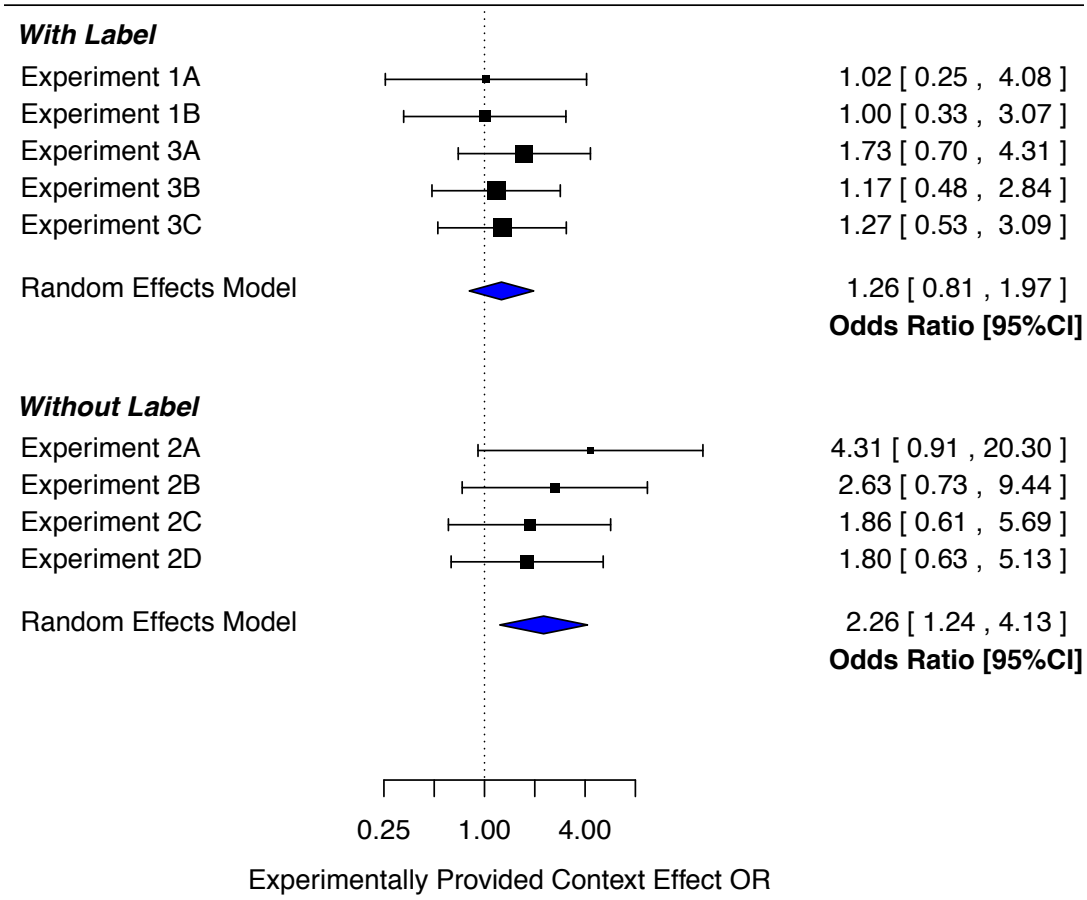

Figure 6.5. Random effects meta-analysis on the effect of experimentally provided counterfactuals for experiments with (1A-B and 3A-C) and without a labeled zero (2A-D) separately. Odds ratios (OR) greater than 1 indicate an increase in the preference for the M0 Option with pay-context instead of receive-context.

General Discussion

In five experiments, we have shown that participants' choices depend on whether a zero outcome is labeled as "pay £0" or "receive £0", replicating Scholten et al.'s (2016) mutable-zero effect. The odds of choosing an option with a zero component were about double when the zero is described as "pay £0" compared to "receive £0" (Experiments 1A and B and 3A-D).

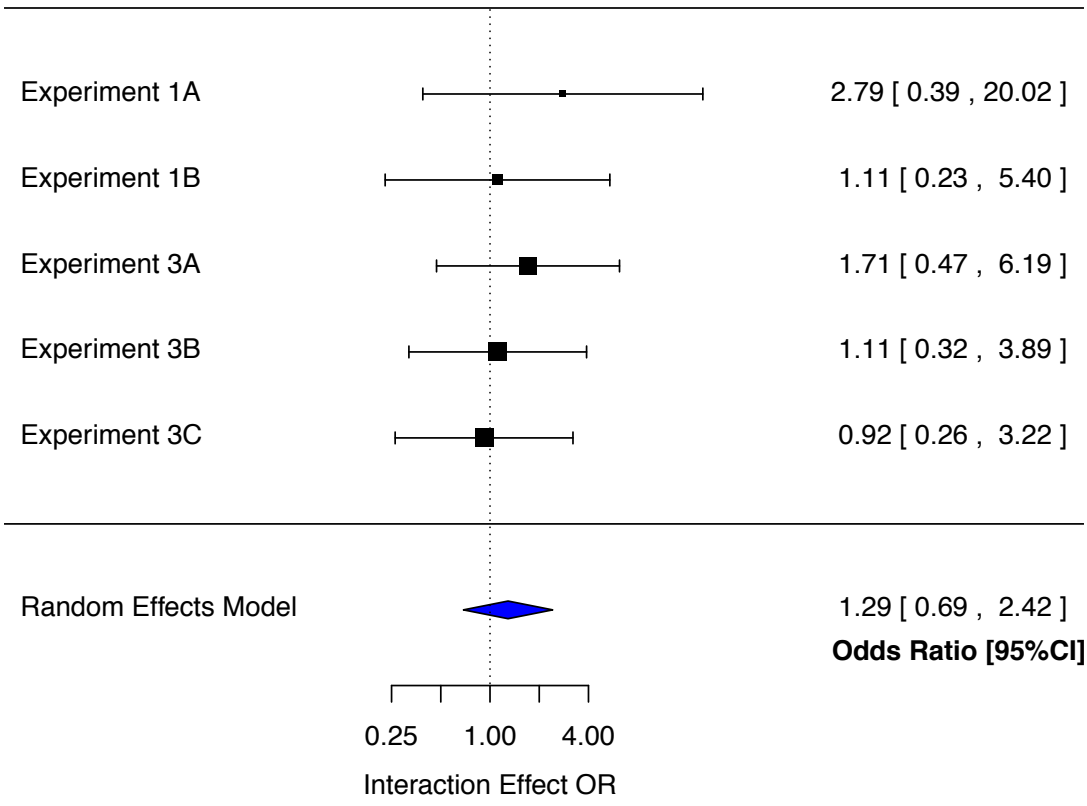

Figure 6.6. Meta-analysis to estimate the difference of the effect of experimentally provided counterfactuals between the pay-category and the receive-category. Odds ratios greater than 1 indicate that the difference between receive-context and pay-context with a "receive zero" attribute is bigger than the difference between receive-context and pay-context with a "pay zero" attribute.

We also found that a zero presented alone as just "£0", without "pay" or "receive" labels, is affected by values in the experimental context (Experiments 2A-D), exhibiting a typical context-effect in the evaluation of magnitudes. But, despite this result and a large body of experiments showing that values in the experimental context affect choice (e.g., Boyce, Brown, & Moore, 2010; Birnbaum, 1992; Carmon & Ariely, 2000; see Stewart et al., 2015, for a recent review), counterfactual values provided in the experimental context do not seem to affect the mutable-zero effect in which a zero is

accompanied by a "pay" or "receive" label. For example, "£0" appears good in the context of "pay £100", "pay £200" and "pay £300" and bad in the context of "receive £100", "receive £200", and "receive £300", but "pay £0" and "receive £0" are unaffected by the same context. This is a quite remarkable finding: just adding a single word "pay" or "receive" seems to have completely blocked, or at least greatly attenuated, effects of context which are well-replicated across many laboratories.

In the Introduction we described three hypotheses about how the mutable-zero effect might work. The first was the Any-Counterfactuals Hypothesis, in which the mutable-zero effect works because the mutable zero leads people to generate counterfactuals, and these counterfactuals work in just the same way as any other counterfactuals. By comparing the size of the effect of the labels "pay" or "receive" in the mutable-zero effect with the size of the effect of experimentally provided counterfactuals when the zero is unlabeled, we see our estimates of the effect sizes are similar, which might suggest a similar mechanism. But, as described above, the experimentally provided counterfactuals presented with a mutable zero do not have an effect. This suggests that the experimentally provided counterfactuals are not being treated in the same way as the counterfactuals that we hypothesize the labels "pay" and "receive" cause people to generate.

The second hypothesis was the Category-Consistent-Counterfactuals Hypothesis, in which counterfactuals only have an effect when they match the sign of the zero. So here, we only expect the counterfactuals of "pay £100", "pay £200" and "pay £300" to have an effect when the zero is labeled with "pay". And we would only expect counterfactuals of "receive £100", "receive £200", and "receive £300" to have an effect

when the zero is labeled with “receive”. But, as described above, we did not find an effect of the experimentally provided counterfactuals, even when the labels match that of the zero. For the Category-Consistent-Counterfactuals Hypothesis to work, we need to assume that the counterfactuals people generate match the magnitudes of the experimentally provided counterfactuals. So we would have to assume that people did compare “pay £0” with counterfactuals “pay £100”, “pay £200”, and “pay £300”. But in “receive £0” condition with “pay £100”, “pay £200”, and “pay £300” we would need to assume people to disregard the “pay” counterfactuals, and replace them with generated counterfactuals of about “receive £100”, “receive £200”, and “receive £300”. In our experiments where all the sums are round hundreds less than £500, assuming people generate counterfactuals like “receive £100” and “receive £300” when presented with “receive £0” seems quite plausible. Because these generated counterfactuals are about the same size as the experimentally provided counterfactuals, they end up having the same effect. In short, if people ignore counterfactuals from the opposite category and instead generate counterfactuals of about the same magnitude from the same category, we’d see no effect of changing the experimentally provided counterfactuals.

Thus, the Category-Consistent-Counterfactual Hypothesis could be consistent with the data, where gains are only compared against other gains and losses are only compared against other losses (McGraw et al., 2010). Walasek and Stewart (2015) found evidence consistent with this suggestion. In their experiment, people were offered the chance to play or reject 50/50 lotteries with a gain and a loss. For example, would you play a 50/50 lottery offering a gain of $10 and a loss of $10? In a condition where the maximum gain on offer in the experiment was $40 and the maximum loss on offer in the

experiment was $20, Walasek and Stewart found classic loss-averse behavior with people rejecting the gain $10 lose $10 lottery. But with the ranges of gains and losses on offer reversed, so that the maximum gain was $20 and the maximum loss was $40, people were much more likely to accept the gain $10 lose $10 lottery. If people were simply evaluating all magnitudes, without segregating them into gains and losses, this would not occur, because the magnitude of amounts on offer is the same in both situations (all between $0 and $40). Instead people seem to evaluate a gain against only the range of gains and a loss against only the range of losses.

The suggestion that people only make comparisons within the same category is related to older ideas about the alignability of attributes. For example, Slovic and MacPhillamy (1974) found that when participants compared pairs of students, they focused on dimensions that the pair had in common. In similarity, differences on alignable attributes are weighted more heavily and this alignment is relevant in decision processes (Medin, Goldstone, & Markman, 1995; Zhang & Markman, 2001). In valuation and choice, alignable dimensions are more evaluable and weight more heavily in judgments and decisions (see the evaluability hypothesis, Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999; see Hsee & Zhang, 2010, for a review).

In the third, Unhappiness-Induced-Counterfactuals Hypothesis, people only make comparisons with the experimentally provided counterfactuals in the “receive zero” conditions because, unlike “pay zero”, the mutable zero counterfactuals generated by participants, lead “receive zero” to be evaluated negatively. We found little evidence for the Unhappiness-Induced-Counterfactuals Hypothesis, under which we expected an interaction between the mutable zero category and the experimentally provided

counterfactuals. We don't doubt Kassam et al.'s finding that initial negative evaluations are more likely to lead to further comparison, because we have replicated their findings in several experiments (see Chapter 5). One reason why there was no interaction effect here between the experimentally provided counterfactuals and the mutable zero category could be that making satisfying comparisons may be a very important mechanism in judgment (the dependent variable in Kassam et al.'s studies), but could be less important in choice (the dependent variable in our studies). We don't know if this was the key difference between Kassam et al.'s experiment and ours. Another explanation for the mutable-zero effect could have nothing to do with counterfactuals and comparison processes. The effect could arise through the associations we make when we are presented with "pay zero", which could be associated with positive outcomes, and "receive zero", which could be associated with negative outcomes. That is, the word "pay" would acquire a negative associative strength by repeated pairing with bad outcomes and the word "receive" would acquire positive associative strength by repeated pairings with good outcomes (as in Hebbian theory, Hebb, 1949; see Shanks 2006, 2010, for recent reviews). However, this associative learning account cannot explain why the experiment provided counterfactuals have an effect with no "pay" or "receive" accompanying a zero, but then have no effect for the mutable "pay zero" and "receive zero".

We have replicated the mutable zero effect, in which an option with a zero outcome is more attractive when the outcome is described as "pay 0" instead of "receive 0". We have also found that although other foregone values provide counterfactuals against which an unlabeled "0" is compared, including a label "pay" or "receive" disrupts

this comparison. We think this is because people only make within-payment or within-receipt comparisons, and do not compare across categories.

7 Conclusion

Here, I will give an overview of the results and conclusions from previous chapters and then propose theoretical and methodological implications of our findings. The aim of this thesis was to test potential mechanisms involved in choice and judgment. We used the DbS model to make predictions and investigate whether people choose options and evaluate outcomes according to their rank position within an underlying sample and under which circumstances comparison processes are involved in risky choice and affective valuation. We thereby test for the DbS proposed cognitive mechanisms of pairwise comparisons and the role of working memory in choice valuation.

Chapter 2

Overview. Chapter 2 tested the robustness and the DbS interpretation of the SRH effect. The SRH effect describes the dependence of the shape of the preference functions on the distribution of attribute values used in the choices presented to participants. In a four-level analysis, we reanalyzed already existing data sets (Level 1), we replicated the experiments with varying experimental properties (Level 2) and extended the experiments from a between-subjects design to a within-subjects design, where participants unbeknownst to them were presented with attributes originating from two distributions (Level 3). Level 3 tested whether the SRH effect is caused by the DbS mechanism of sampling, pairwise comparisons between the current options and other available options from memory and frequency accumulation. If indeed the combination of these cognitive tools is what causes the SRH effect, then exposing participants to (in their world) one distribution should not lead to differences in preference functions when choice data is split retrospectively for the analysis. In contrast to the DbS prediction, the

SRH effect still emerges. In Level 4, the effect size comparison between the different designs via meta-analysis lets us conclude that the DbS mechanisms can, at best, only be part of the explanation for the effect. We do not know of any model that can account for the differences in the preference functions generated through the differently skewed distributions of attribute values presented to participants in the non-flagged versions of our experiments. However, our findings also hint to problems in the modelling procedure, which are probably causing the SRH effect (see Stewart, Canic, & Mullett, 2016). Are other results that found differences between conditions affected by these findings?

Implications. The findings of this chapter raise important methodological aspects: While the call for replications (Open Science Collaboration, 2015) is indeed crucial and an important step in the right direction, here we find that problems with the decision by sampling account only arise when we went beyond replication. Moreover, by using Levitt and List's (2009) methodology and our subsequent meta-analysis, it made us realize that there is a problem with the conclusions that might be drawn when only using exact replications to confirm certain findings. Specifically, we now know that the SRH effect reported in Stewart, Reimers and Harris (2015) cannot be attributed to DbS proposed mechanism. Although the DbS model predicted the SRH results in advance, and the SRH result replicates well, our extensions show that DbS is almost certainly not the explanation of the effect. Therefore, the SRH effect must be caused either via a yet unspecified mechanism. Alternatively, although our modeling procedure of fitting expected utility to reveal risk aversion is pretty standard in this field, the results here suggest this could be the problem, and indeed it is (Stewart, Canic, & Mullett, 2016). There is no reason to believe that other results using the same or a similar procedure do

not suffer from the same draw-backs. Furthermore, these results prompted more specific investigations in identifying the cause for the SRH effect. As our findings show, the effect is robust and substantial, despite the fact that DbS is unlikely to be its origin. This fact further questions preference-based accounts of choice.

Chapter 3

Overview. The Chapter 3 experiments are set up very similarly to the Chapter 2 experiments. We were still interested in the choice mechanism responsible for the SRH effect. Our premise was that cognitive load interferes with the decision making process under risk (Unsworth & Heitz, 2007). The DbS proposed cognitive tools are sampling, pairwise comparisons and frequency accumulation, and they should all depend on cognitive capacity and thus might all be affected by cognitive load. Therefore, working memory is assumed to play a major role in the proposed mechanism. Specifically, DbS predicts that whilst under cognitive load, it would be hard for participants to either keep track of the sample of attribute values presented, or the probability of making mistakes could be increased during pairwise comparisons or frequency accumulation, if this is how the process works. We used a blocked within-subjects design where participants made consecutive choices, sometimes under high cognitive load and sometimes under low or no cognitive load. The estimated value functions resulting from choices under high cognitive load, closely resembled the value functions resulting from choices under low or no cognitive load. Does this mean that choices remained unaffected by cognitive load? We tested additional hypotheses, namely that cognitive load leads to noisier choices (Andersson et al., 2013) and that cognitive load leads to more risk averse choices (Benjamin, Brown, & Shapiro, 2013). Using the same modelling procedure, we found no

effect of cognitive load on risk aversion. However, cognitive load did decrease choice consistency across all three experiments run in this chapter.

Implications. To our knowledge, this was the first attempt at using a cognitive load manipulation in order to detect possible mechanisms of the SRH effect. We took advantage of our design to check for evidence of other hypotheses also, namely that cognitive load increases risk aversion and that cognitive load decreases choice consistency. Whilst the SRH effect seemed unaffected by cognitive load, which further points away from DbS as its originating mechanism, our estimates of decreasing choice consistency confirm Andersson et al.'s findings (2013). Especially given that risk aversion remained unaffected by cognitive load, our findings also suggest that studies relating cognitive capacities and cognitive load with risky decision making need to be able to control for choice consistency. Otherwise, spurious results cannot be excluded. It is also worth thinking about other results that might be affected by the same issue. As more evidence emerges that risk aversion remains unaffected by cognitive load (see Olschewski, Scheibehenne, & Rieskamp, 2015), it raises the question whether results of studies which find a relationship with cognitive capacities and choice behavior so far (and which we reviewed earlier), are confounded.

Chapters 4 and 5

Overview. In Chapter 4 we review evidence for the rank hypothesis and estimate rank effects in judgment and risky choice with yet unpublished data collected by Stewart and Reimers (2008). We present over 20 papers with different response modes and distribution manipulations with varying dependent variables including behavioral or neurophysiological measures. The findings support the rank hypothesis and none of the

studies uses modelling procedures that might be contaminated with the same problems that the previous two chapters face. Some studies found context effects when manipulating the range or the skew of the response scale (Birnbaum, 1992; Stewart et al., 2003), other manipulated the range or skew of the attributes in the choice set (Vlaev et al., 2007; Walasek & Stewart, 2015) and still others manipulate the background context (Janiszewski & Lichtenstein, 1999). There is not just behavioural, but also neurophysiological evidence for context effects; Studies found that brain regions associated with valuation, are activated according to the relative value of prizes on offer (Mullett & Tunney, 2013), and other activation pointed to non-linear mapping between objective reward and subjective value (Rigoli et al. 2016). There is also evidence for comparison processes when looking at the monkey brain in more detail, by measuring the activation in single cells: Activation was higher for an outcome that was favored over another outcome, than if the same outcome was not favored over an alternative outcome, showing relativity in neural responding. We present an additional study that is in line with the reviewed findings. In seven experiments, Stewart and Reimers (2008) asked participants to rate the attractiveness of prospects or to choose between two prospects. They varied the rank position of one or both attribute value (either the probability only or the amount only or both) between conditions and found higher attractiveness ratings for common prospects if these prospects were presented with positively skewed attribute values than if they were presented with negatively skewed attribute values. This is because a common value will have a higher rank position in a positively skewed distribution than a negatively skewed distribution, other factors like range held constant. They also found that the preference for a critical prospect reversed if its attribute values

ranked lower than if they ranked higher within the attribute values presented up until the critical choice. Given the SRH effect in the within-subjects design experiments and no effect of cognitive load in Chapter 3, we cannot conclude that the malleability of preference functions can be attributed to DbS proposed cognitive tools. Here, however, the context effects can only emerge through adaptation to the attribute values the participants were exposed to during the course of the experiment. Thus even if the estimated value functions do not originate via a decision-by-sampling-like process, the conclusion remains that attributes of prospects are evaluated relative to other available attributes in the environment and or in memory.

In Chapter 5 we build on the findings from Chapter 4. We know that prospects with attribute values that rank higher among other attribute values in the sample are also perceived as more attractive than prospects with attribute values that rank lower (Stewart & Reimers, 2008). We combine this finding with findings from Kassam et al. (2011): They asked participants to choose between two scratch panels, which hid different prizes. They found prize sensitivity in positive affect ratings when participants win the lower of the two prizes, but insensitivity when participants win the higher of the two prizes. Thus, participants report an equally high positive affect no matter if they win $7 instead of $5 or $3 instead of $1. However, participants report a lower positive affect when they win $1 instead of $3 than when they win $5 instead of $7. Kassam et al. attribute this asymmetry to the comparisons that people make. They test this hypothesis by introducing cognitive load and find evidence that supports their hypothesis: Under cognitive load–presumably unable to make further comparisons–participants' positive affect ratings were suddenly insensitive to the absolute prize value. Only under low cognitive load did prize

sensitivity remain. We use Kassam et al.'s insight and test the hypothesis that people compare their prize not just with the salient foregone prize that remains on the screen, but also with other recently experienced prizes. We manipulated the rank position of a won and salient foregone prize and expected to find lower positive affect ratings when the prize won was several ranks lower than the foregone prize and relatively high affect ratings when the prize won was only one rank lower than the foregone prize. In contrast to our expectations and the rank hypothesis, participants positive affect ratings seem unaffected by the rank position of the prize and its alternative within the sample of previously experienced prizes. In contrast to Kassam et al.'s findings, we found no effect of cognitive load on the reported positive affect. However, we might not even expect an effect, given that we did not find evidence for the rank hypothesis either. So when do people form judgments through comparison processes and adjust to the context and when do they not adjust?

Implications. We establish a firm basis for the relativity of prospects, outcomes and their context-dependent representation in the brain, and we were surprised by the null results in Chapter 5. They imply that when a prize (outcome in the domain of gains) is evaluated in the context of a higher prize, the relative rank position of both prizes within the experimental context is not used to evaluate the prize. We could speculate that the null result is due to the prize range being very familiar, which made the context manipulation too weak, or that the fact that the prizes were hypothetical made participants not engage enough with the experiment. However, whilst one could explore other prize ranges or use incentives to be able to exclude those reasons for sure, it would not explain why some findings in Chapter 4 still display rank effects even though they

were conducted under similar conditions. Still, it might be as simple as the familiarity argument (see Mellers, Ordonez and Birnbaum, 1992): Because people are very familiar with the feeling of having received some amount that ranges between £1 and £100, the comparisons that automatically come to people's minds are numbers from £1 to £100. This would make the comparisons with the prizes from the experimental context some among many others, which in turn would be insufficient to yield observable rank effects. If this is true, one should establish when and to what degree long-term memory as well as effects from the immediate context influence the evaluation of outcomes.

Chapter 6

Overview. In Chapter 6 we used the mutable-zero effect (Scholten et al., 2016) to look at which counterfactuals influence the evaluation of an option. The mutable-zero effect is the phenomenon that an option is more likely to be chosen when it entails a zero outcome that is described as "pay zero" instead of "receive zero". Scholten et al. hypothesized that the effect emerges through the comparisons that people make: When people encounter "pay zero", other payments come to mind, which make "pay zero" comparably desirable. On the other hand, when people encounter "receive zero", other receipts come to mind, which make "receive zero" comparably undesirable. We tested whether all counterfactuals provided by the experimental context can shift preference, or whether only counterfactuals provided by the experimental context that match the mutable-zero category (category-consistent counterfactuals) can shift preference. We also tested an interaction hypothesis, in which only when people encounter a disappointing "receive zero" do people make further comparisons so that all counterfactuals will be able to shift participants' preferences. When participants encounter a happy "pay zero",

participants will not make further comparisons so that the counterfactuals provided by the experimental context will have no effect. We consistently found that whilst experimentally provided context influences valuation when zero stands on its own without "pay" or "receive" standing in front of it, counterfactuals provided by the context affect valuation in the same way as "pay zero" or "receive zero" do. However, as soon as zero is preceded by "pay" or "receive" the category blocks comparisons with counterfactuals that are inconsistent with the mutable-zero category so that only category-consistent counterfactuals are decisive for the evaluation process. This is a quite striking result: Adding a single word (either "pay" or "receive") completely eliminated the otherwise strong effects of the distribution of attribute values in the experiment.

Implications. The findings reported in this chapter nicely build up on the hypothesis that long-term memory as well as the immediate context play important roles in the evaluation process of alternatives. Here, the long-term memory component is represented in terms of a representation of amounts that belong to the category "pay" or "receive". Even if the experimental or the choice context provide actual counterfactuals, they are not taken into account when they are not consistent with the category of the attribute to be evaluated. Thus an existing category will block category-inconsistent counterfactual values. This also suggests some kind of monitoring process and representation of the world that involves both, a long-term and a working-memory component.

Conclusion

We have seen that the manipulation of the choices presented to participants does robustly alter the utility functions elicited from those participants. But the results are

unlikely to be due to the decision-by-sampling mechanism in which the attribute values in the choices are evaluated by accumulating favorable pairwise ordinal comparisons between them. We reach this conclusion because we find different utility functions are elicited when different subsets of retrospectively partitioned choice sets are used to estimate them. This points to a problem using models like prospect theory and expected utility to reveal utility functions (and which we confirm in Stewart, Canic, & Mullett, 2016).

We have also not found any evidence that working memory plays a role in decision making; not in the SRH effect nor in the "winners love winning losers love money" effect, beyond affecting the consistency of people's choices: Higher load makes people a bit more random. However, given that there might be an issue with the estimation procedure in the SRH effect and that we did not find a robust effect of the experimental context in the "winners love winning losers love money" effect, we should perhaps not conclude too much from this failure.

There is overwhelming evidence for context effects like those that decision by sampling predicts, and we have used a meta-analysis on some simpler experimental designs where we can see shits in ratings for common gambles, or shifts in preference for common choices, as we vary the distribution of attribute values across experimental conditions. We have also been able to place boundaries of what kinds of comparisons matter in affecting valuations, and in exploring the mutable zero effect, we have added to evidence that comparisons only spontaneously happen within gains, or within losses, but not across gains and losses.

References

Abele, A. (1985). Thinking about thinking: Causal, evaluative and finalistic cognitions about social situations. *European Journal of Social Psychology, 15*, 315–332.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60.

Adaval, R., & Monroe, K. B. (2002). Automatic construction and use of contextual information for product and price evaluations. *Journal of Consumer Research*, *28*, 572–588.

Aldrovandi, S., Wood, A. M., & Brown, G. D. (2013). Sentencing, severity, and social norms: A rank-based model of contextual influence on judgments of crimes and punishments. *Acta psychologica*, *144*, 538–547.

Alempaki, D., Canic, E. Mullett, T., Tufano, F., Matthews, W., Stewart, N., Starmer, C. (2016). *Examining how utility and weighting function get their shapes: A multi-level, quasi-adversarial replication.* Manuscript submitted for publication.

Allais, M. (1953). L'extension des théories de l'équilibre économique général et du rendement social au cas du risque. *Econometrica, Journal of the Econometric Society*, 269–290.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396-408.

Andersson, O., Tyran, J. R., Wengström, E., & Holm, H. J. (2013). *Risk aversion relates to cognitive ability: Fact or fiction?* (IFN Working Paper No.964). Stockholm: Research Institute of Industrial Economics.

Arvai, J., Gregory, R., Ohlson, D., Blackwell, B., & Gray, R. (2006). Letdowns, wake-up calls, and constructed preferences: people's responses to fuel and wildfire risks. *Journal of Forestry, 104,* 173–181.

Ayal, S., & Hochman, G. U. Y. (2009). Ignorance or integration: The cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making*, *22*, 455–474.

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. *Handbook of the Economics of Finance*, *1*, 1053–1128.

Baron, J. (2007). Thinking and deciding (4th ed.). Cambridge, England: Cambridge University Press.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review, 118,* 175–192.

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*, 1561–1580.

Beauchamp, J. P., Benjamin, D. J., Chabris, C. F., & Laibson, D. I. (2012). *How malleable are risk preferences and loss aversion*. Working paper, Harvard University, Cambridge, MA. http://economics. cornell. edu/dbenjamin/HowMalleableareRisk PreferencesandLossAversion_3-10-2012-Combined. pdf.

Benartzi, S., & Thaler, R.H. (2001). Naive diversification strategies in defined contribution saving plans. *American Economic Review. 91*, 79–98.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, *11*, 1231–1255.

Bergman, O., Ellingsen, T., Johannesson, M., & Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters*, *107*, 66–68.

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

Birnbaum, M. H. (1992). Violations of the monotonicity and contextual effects in choice-based certainty equivalents. *Psychological Science, 3,* 310–314.

Birnbaum, M. H. (2004). Causes of Allais common consequence paradoxes: An experimental dissection. *Journal of Mathematical Psychology*, *48*, 87–106.

Birnbaum, M. H. (2008a). Evaluation of the priority heuristic as a descriptive model of risky decision making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115,* 253–260.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*, 463–501.

Birnbaum, M. H. (2010). Testing lexicographic semiorders as models of decision making: Priority dominance, integration, interaction, and transitivity. *Journal of Mathematical Psychology, 54*, 363–386.

Birnbaum, M. H., & Bahra, J. P. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science*, *53*, 1016–1028.

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes*, *67*, 91–110.

Birnbaum, M. H., & Navarrete, J. B. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, *17*, 49–79.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, *37*, 48–74.

Bohm, P., Linden, J., & Sonnegard, J. (1997). Eliciting Reservation Prices: Becker DeGroot-Marschak Mechanisms vs. Markets. *Economic Journal, 107*, 1079–1089.

Boyce, C. J., Brown, G. D., & Moore, S. C. (2010). Money and happiness rank of income, not income, affects life satisfaction. *Psychological Science*, *21*, 471–475.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*, 409–432.

Brown, G. D., Gardner, J., Oswald, A. J., & Qian, J. (2008). Does Wage Rank Affect Employees' Well-being? *Industrial Relations: A Journal of Economy and Society*, *47*, 355–389.

Bruza, B., Welsh, M. B., Navarro, D. J. (2008). Does memory mediate susceptibility to cognitive biases? Implications of the Decision-by-Sampling theory. *Proceedings of the 30th Annual meeting of the Cognitive Science Society,* 1498–1503.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2009). Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, *106*, 7745–7750.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459.

Camerer, C. F. (1992). Recent tests of generalizations of expected utility theory. In W. Edwards (Ed.), *Utility theories: Measurements and applications* (pp. 207–251). Boston: Kluwer Academic Publishers.

Canic, E. (2016). *Affective evaluation of monetary outcomes is unaffected by rank position.* (In unpublished doctoral thesis). University of Warwick, Coventry, UK.

Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequality at work: The effect of peer salaries on job satisfaction. *The American Economic Review*, *102*, 2981–3003.

Carmon, Z., & Ariely, D. (2000). Focusing on the forgone: How value can appear so different to buyers and sellers. *Journal of Consumer Research, 27,* 360–370.

Clark, A.E., Frijters, P., & Shields, M.A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin Paradox and other puzzles. *Journal of Economic Literature*, *46*, 95–144.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*, 20–33.

Colom, R., Flores-Mendoza, C., & Rebollo, I. (2003). Working memory and intelligence. *Personality and Individual Differences*, *34*, 33–39.

Conway, A. R., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*, 547–552.

Corbin, J., McElroy, T., & Black, C. (2010). Memory reflected in our decisions: Higher working memory capacity predicts greater bias in risky choice. *Judgment and Decision Making, 5*, 110–115.

Corrigan, J. R., Drichoutis, A. C., Lusk, J. L., Nayga, R. M., & Rousu, M. C. (2011). Repeated rounds with price feedback in experimental auction valuation: An adversarial collaboration. *American Journal of Agricultural Economics*, aar066.

Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, *78*, 97–119.

De Neys, W. (2006). Dual processing in reasoning two systems but one reasoner. *Psychological science*, *17*, 428–433.

Dhar, R., Nowlis, S. M., & Sherman, S. J. (2000). Trying hard or hardly trying: An analysis of context effects in choice. *Journal of Consumer Psychology*, *9*, 189–200.

Dohmen, T. J., Falk, A., Huffman, D., & Sunde, U. (2007). *Are risk aversion and impatience related to cognitive ability?* IZA Discussion Papers, No. 2735, http://nbn-

resolving.de/urn:nbn:de:101:1-20080402396

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory process model for judgments of likelihood. *Psychological Review, 106,* 180–209.

Dougherty, M. P. R., & Hunter, J. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113,* 263–282.

Dougherty, M. P. R., & Hunter, J. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31,* 968–982.

Drichoutis, A. C., & Nayga, R. M. (2013). Eliciting risk and time preferences under induced mood states. *The Journal of Socio-Economics*, *45*, 18-27.

Dunlop, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177.

Dunning, D., & Hayes, A.F. (1996). Evidence for egocentric comparison in social judgment. *Journal of Personality and Social Psychology*, *71*, 213–229.

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology*, *50*, 201–214.

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review*, *69*, 109–135.

Elliott, R., Agnew, Z., & Deakin, J. F. W. (2008). Medial orbitofrontal cortex codes relative rather than absolute value of financial rewards in humans. *European Journal of Neuroscience*, *27*, 2213–2218.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in Psychological Science*, *11*, 19–23.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128,* 309–331.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*, 168–192.

Ert, E., & Erev, I. (2013). On the descriptive value of loss aversion in decisions under risk: Six clarifications. *Judgment and Decision Making*, *8*, 214–235.

Etchart-Vincent, N. (2004). Is probability weighting sensitive to the magnitude of consequences? An experimental investigation on losses. *Journal of Risk and Uncertainty*, *28*, 217–235.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Reviews in Psychology*, *59*, 255–278.

Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, *40*, 147–180.

Fehr-Duda, H., Epper, T., Bruhin, A., & Schubert, R. (2011). Risk and rationality: The effects of mood and decision rules on probability weighting. *Journal of Economic Behavior & Organization*, *78*, 14–24.

Fiedler, K. (2010). How to study cognitive decision algorithms: The case of the priority heuristic. *Judgment and Decision Making*, *5*, 21–32.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178.

Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 580–597.

Franco-Watkins, A. M., Rickard, T. C., & Pashler, H. (2015). Taxing executive processes does not necessarily increase impulsive decision making. *Experimental Psychology*, *57,* 193–201.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42.

Friedman, D., Isaac, R. M., James, D., and Sunder, S. (2014). *Risky curves. On the empirical failure of expected utility*. Routledge, NY: Princeton University Press.

Galanter, E., & Pliner, P. (1974). *Cross-modality matching of money against other continua* (pp. 65-76). Springer Netherlands.

Getz, S. J. (2013). *Cognitive Control and Intertemporal Choice: The Role of Cognitive Control in Impulsive Decision Making* (Unpublished doctoral thesis). Princeton University, Princeton, NJ.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3-34). New York: Oxford University Press.

Gilbert, D.T., Giesler, R.B., & Morris, K.A. (1995). When comparisons arise. *Journal of Personality and Social Psychology*, *69*, 227–236.

Gilbert, D.T., Lieberman, M.D., Morewedge, C.K., & Wilson, T.D. (2004). The peculiar longevity of things not so bad. *Psychological Science*, *15*, 14–19.

Glöckner, A., & Betsch, T. (2008). Do people make decisions under risk based on ignorance? an empirical test of the priority heuristic against cumulative prospect theory. *Organizational Behavior and Human Decision Processes, 107,* 75–95.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association, 1,* 114-125.

Haggag, K., & Paci, G. (2014). Default tips. *American Economic Journal: Applied Economics*, *6*, 1–19.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Helson, H. (1964). Current trends and issues in adaptation-level theory. *American Psychologist*, *19*, 26–38.

Hilbig, B. E. (2008). One-reason decision making in risky choice? A closer look at the priority heuristic. *Judgment and Decision Making*, *3*, 457–462.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 29,* 298–306.

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

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

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

Holroyd, C. B., Larsen, J. T., & Cohen, J. D. (2004). Context dependence of the event-related brain potential associated with reward and punishment. *Psychophysiology*, *41*, 245–253.

Jameson, T. L., Hinson, J. M., & Whitney, P. (2004). Components of working memory and somatic markers in decision making. *Psychonomic Bulletin & Review, 11,* 515–520.

Janiszewski, C., & Lichtenstein, D. R. (1999). A range theory account of price perception. *Journal of Consumer Research*, *25*, 353–368.

Janiszewski, C., & Van Osselaer, S. M. (2000). A connectionist model of brand-quality associations. *Journal of Marketing Research*, *37*, 331–350.

Jensen, A. R. (1998). The g factor: The science of mental ability. *Politics and the Life Sciences*, *17,* 230–232.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: a query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 461–474.

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115,* 263–272.

Kahneman, D. (2003a). Experiences of collaborative research. *American Psychologist*, *58*, 723–730.

Kahneman, D. (2003b). Maps of bounded rationality: Psychology for behavioral economics. *The American Economic Review*, *93*, 1449–1475.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Frederick, S. (2007). Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Sciences*, *11*, 45–46.

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

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society, 47,* 263–291.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 66–71.

Kanouse, D. E., & Hanson Jr, L. R. (1987). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 47–62). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kassam, K. S., Morewedge, C. K., Gilbert, D. T., & Wilson, T. D. (2011). Winners love winning and losers love money. *Psychological Science, 22,* 602–606.

Kobayashi, S., de Carvalho, O. P., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *The Journal of Neuroscience*, *30*, 534–544.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics, 121,* 1133–1165.

Kyllonen, P. C. (2002). g: Knowledge, speed, strategies, or working-memory capacity? A systems perspective. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *The General Factor of Intelligence: How General Is It* (pp. 415-445). Mahwah, NJ: Lawrence Erlbaum Associates.

Laming, D. R. J. (1997). The measurement of sensation. London: Oxford University Press.

Larsen, J. T., McGraw, A. P., Mellers, B. A., & Cacioppo, J. T. (2004). The agony of victory and thrill of defeat mixed emotional reactions to disappointing wins and relieving losses. *Psychological Science*, *15*, 325–330.

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*, 753–772.

Levitt, S. D., & List, J. A. (2009). Field experiments in economics: the past, the present, and the future. *European Economic Review*, *53*, 1–18.

Lewicka, M., Czapinski, J., & Peeters, G. (1992). Positive-negative asymmetry or 'When the heart needs a reason'. *European Journal of Social Psychology*, *22*, 425–434.

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

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*, 805–824.

Loomes, G. (2010). Modeling choice and valuation in decision experiments. *Psychological Review, 117*, 902–924.

Luce, R.D. (1959) *Individual Choice Behavior.* New York: John Wiley & Sons.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, *143*, 1185–1198.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *The American Economic Review*, *104*, 277–290.

Mazar, N., Koszegi, B., & Ariely, D. (2014). True Context-dependent Preferences? The Causes of Market-dependent Valuations. *Journal of Behavioral Decision Making*, *27*, 200–208.

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

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision-processes. *Psychonomic Bulletin & Review, 2,* 1–19.

Medvec, V. H., Madey, S. F., & Gilovich, T. (1995). When less is more: counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality and Social Psychology*, *69*, 603–610.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

Mellers, B. A., Ordoñez, L. D., & Birnbaum, M. H. (1992). A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes*, *52*, 331–369.

Mellers, B.A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*, 210–214.

Melrose, K. L., Brown, G. D., & Wood, A. M. (2013). Am I abnormal? Relative rank and social norm effects in judgments of anxiety and depression symptom severity. *Journal of Behavioral Decision Making*, *26*, 174–184.

Morewedge, C.K., Gilbert, D.T., Myrseth, K.O.R., Kassam, K.S., & Wilson, T.D. (2010). Consuming experience: Why affective forecasters overestimate comparative value. *Journal of Experimental Social Psychology*, *46*, 986–992.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence–their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 61–65.

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity–facets of a cognitive ability construct. *Personality and Individual Differences*, *29*, 10171045.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*, 147–152.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*, 366–387.

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

Payne, J. W., Bettman J. R., & Schkade D.A. (1999). Measuring constructed preferences: towards a building code. *Journal of Risk and Uncertainty, 19,* 243–270.

Payne, J. W., Samper, A., Bettman, J. R., & Luce, M. F. (2008). Boundary conditions on unconscious thought in complex decision making. *Psychological Science*, *19*, 1118–1123.

Parducci, A. (1959). An adaptation-level analysis of ordinal effects in judgments. *Journal of Experimental Psychology*, *58*, 239–246.

Parducci, A. (1963). Range-frequency compromise in judgment. *Psychological Monographs: General and Applied*, *77*, 1–50.

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

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In L. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 127–141). New York: Academic Press.

Parducci, A., Calfee, R. C., Marshall, L. M., & Davidson, L. P. (1960). Context effects in judgment: Adaptation level as a function of the mean, midpoint, and median of the stimuli. *Journal of Experimental Psychology*, *60*, 65–77.

Peeters, G. (1971). The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology, 1*, 455–474

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*, 33–60.

Prelec, D., Wernerfelt, B., & Zettelmeyer, F. (1997). The role of inference in context effects: Inferring what you want from what is available. *Journal of Consumer Research*, *24*, 118–126.

Pyszczynski, T., Greenberg, J., & LaPrelle, J. (1985). Social comparison after success and failure: Biased search for information consistent with a self-serving conclusion. *Journal of Experimental Social Psychology*, *21*, 195–211.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization*, *3*, 323–343.

Rieger, M. O., & Wang, M. (2008). What is behind the priority heuristic? A mathematical analysis and comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*, 274–280.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, *44,* 631–661.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*, 133–148.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion.

Personality and Social Psychology Review, 5, 296–320.

Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical Decision Making, 28,* 850–865.

Rick, S. (2011). Losses, gains, and brains: Neuroeconomics can help to answer open questions about loss aversion. *Journal of Consumer Psychology*, *21*, 453–463.

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nature Communications*, *7*, 12416.

Rigoli, F., Rutledge, R. B., Dayan, P., & Dolan, R. J. (2016). The influence of contextual reward statistics on risk preference. *NeuroImage*, *128*, 74–84.

Sakaguchi, H., & Stweart, N. (2016). *Negative Tweets Have More Reasoning Statements than Positive Tweets: An Evidence for the Negativity Bias in Twitter.* Manuscript in preparation.

Scholten, M., Read, D., Canic, E., & Stewart, N. (2015). *The scope and specificity of the mutable-zero effect.* Manuscript submitted for publication.

Seymour, B., & McClure, S. M. (2008). Anchors, scales and the relative coding of value in the brain. *Current Opinion in Neurobiology*, *18*, 173–178.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*, 546–556.

Shanks, D. R. (2006). Bayesian associative learning. *Trends in Cognitive Sciences, 10,* 477–478.

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61,* 273–301.

Shepard, R.N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrica, 22*, 325–345.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69,* 99–118.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, *29*, 281–295.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Slovic, P., & MacPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance, 11,* 172–194.

Sprenger, A., & Dougherty, M. R. (2006). Differences between probability and frequency judgments: The role of individual differences in working memory capacity. *Organizational Behavior and Human Decision Processes*, *99*, 202–211.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*, 701–717.

Starmer, C., & Sugden, R. (1989). Violations of the independence axiom in common ratio problems: An experimental test of some competing hypotheses. *Annals of Operations Research*, *19*, 79–102.

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

Stewart, N., Canic, E., & Mullett, T. (2017). *On the futility of estimating utility functions: Why the parameters we measure are wrong, and why they do not generalize.* Manuscript submitted for publication.

Stewart, N., Chater, N., & Brown, G.D.A. (2006). Decision by sampling. *Cognitive Psychology, 53,* 1–26.

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General, 132,* 23–46.

Stewart, N. & Reimers, S. (2008). *How the distribution of attribute values influences ratings of risky prospects.* Unpublished manuscript.

Stewart, N., Reimers, S., & Harris, A. J. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*, 687–705.

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, affect, and mode of thinking. *Journal of Personality and Social Psychology*, *49*, 1460–1469.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological Bulletin*, *110*, 67–85.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*, 704–708.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 1039–1061.

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

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science, 22,* 253–260.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence*, *33*, 67–81.

Van Osselaer, S. M., & Alba, J. W. (2000). Consumer learning and brand equity. *Journal of Consumer Research*, *27*, 1–16.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1-48.

Vlaev, I., Chater, N., & Stewart, N. (2007). Financial prospect relativity: context effects in financial decision-making under risk. *Journal of Behavioral Decision Making*, *20*, 273–304.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, *15*, 546–554.

Vlaev, I., Seymour, B., Dolan, R. J., & Chater, N. (2009). The price of pain and the value of suffering. *Psychological Science*, *20*, 309–317.

von Neumann, M., & Morgenstern, O. (1947). *Theory of Games and Economic Behavior* (2nd ed.). Princeton, NJ: Princeton University Press.

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General*, *144*, 7–11.

Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 193–205.

Watkinson, P., Wood, A. M., Lloyd, D., & Brown, G. D. A. (2013). Pain ratings reflect cognitive context: A range frequency model of pain. *Pain, 154,* 743–749.

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice a query-theory account. *Psychological Science*, *18*, 516–523.

Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 767–778.

Whitney, P., Rinehart, C. A., & Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic Bulletin & Review, 15,* 1179–1184.

Wood, A. M., Brown, G. D. A., & Maltby, J. (2011). Thanks, but I'm used to better: A relative rank model of gratitude. *Emotion, 11,* 175–180.

Wu, G., & Markle, A. B. (2008). An empirical test of gain-loss separability in prospect theory. *Management Science*, *54*, 1322–1335.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience*, *24*, 6258–6264.

Yoon, S. O., & Simonson, I. (2008). Choice set configuration as a determinant of preference attribution and strength. *Journal of Consumer Research*, *35*, 324–336.

Zhang, S., & Markman, A. B. (2001). Processing product unique features: Alignability and involvement in preference construction. *Journal of Consumer Psychology*, *11*, 13–27.

Appendix A

The calculation errors in the SRH's original analysis that were discovered as a result of our Level 1 replication analysis are as follows:

- In experiment SRH 1A the original analysis takes into account only 120 instead of 150 questions. However, except from moving the functions slightly, this has no effect on the difference between the two conditions.
- In the description of the analysis of experiment SRH 2B it is mentioned that none of the participants violated dominance in more than 10% of the trials. Our analysis shows that 4 participants violated dominance and therefore should have been excluded from further analysis. Again, the differences between conditions reported by SRH remain significant after this adjustment.

As we explain in the main paper, the Level 1 analysis we apply and report in the manuscript differs from SRH's original analysis in terms of the estimation of both the confidence intervals (i.e., we used a more reliable bootstrapping method) and the revealed functions (i.e., we estimated the revealed functions separately for each condition). To further probe the robustness of the inference from SRH's as well as our own (Level 1) analysis, it is useful to test statistically whether SRH conclusions would still hold when correcting the above calculation errors. Hence, we applied SRH's original analysis (without the refinements we use in the analysis reported in the main paper) to SRH raw data and compared the results with and without taking into account these errors. The results from this comparison are reported in Table A1 below. Even though there are some minor differences in the reported statistics with regard to Experiments SRH 1A and SRH 2B, the results are qualitatively very similar and none of SRH's original claims are affected by correction of these minor errors.

Table A1

Results from the SRH's original analysis using SRH raw data with and without the calculation errors.

	With the calculation errors		**Without the calculation errors**	
Experiment	Differences in concavity in the revealed functions between conditions	Differences in weighting common amounts or common probabilities between conditions	Differences in concavity in the revealed functions between conditions	Differences in weighting common amounts or common probabilities between conditions
SRH 1A	$\chi^2(1)=6.36$, $p=0.012$	$\chi^2(1)=7.05$, $p=0.0079$ for comparison between £200 and £310	$\chi^2(1)= 11.93$, $p=.0006$	$\chi^2(1)=22.75$, $p<0.0001$ for comparison between £200 and £310
SRH 1B	$\chi^2(1)=6.99$, $p=.0082$	$\chi^2(1)=26.96$, $p<0.0001$ for the common £100, $\chi^2(1)=7.16$, $p=0.0074$ for the common £200	$\chi^2(1)=6.98$, $p=.0082$	$\chi^2(1)=27.20$, $p<0.0001$ for the common £100, $\chi^2(1)=6.77$, $p=.009$ for the common £200
SRH 1C	$\chi^2(1)=3.50$, $p=0.06$	$\chi^2(1)=59.79$, $p<0.0001$ for the common £100, $\chi^2(1)=50.47$, $p<0.0001$ for the common £200	$\chi^2(1)=3.49$, $p=.06$	$\chi^2(1)=59.75$, $p<0.0001$ for the common £100, $\chi^2(1)=50.32$, $p<0.0001$ for the common £200
SRH 2A	$\chi^2(1)=2.18$, $p=0.13$	$\chi^2(1)=18.18$, $p<0.0001$ for the common 30%, $\chi^2(1)=14.31$, $p=0.0002$ for the common 70%	$\chi^2(1)=2.18$, $p=0.14$	$\chi^2(1)=18.22$, $p<0.0001$ for the common 30%, $\chi^2(1)=14.41$, $p=0.0001$ for the common 70%
SRH 2B	$\chi^2(1)=181.5$, $p<0.0001$	$\chi^2(1)=41.72$, $p<0.0001$ for the common 50%	$\chi^2(1)=119.6$, $p<0.0001$	$\chi^2(1)=22.8$, $p<0.0001$ for the common 50%

Appendix B

In our Level 3 replication in order to account for the within-subject nature of this series of experiments, we used only one model including all random effects, instead of using separate models for each condition as we did for the between-subject experiments. Figure B1 shows the functional forms and the relative confidence intervals obtained by estimating a separate model for each condition. By comparing Figure B1 to Figure 2.5, it is possible to see that both the estimated functional forms and the confidence intervals are very similar across the two figures.

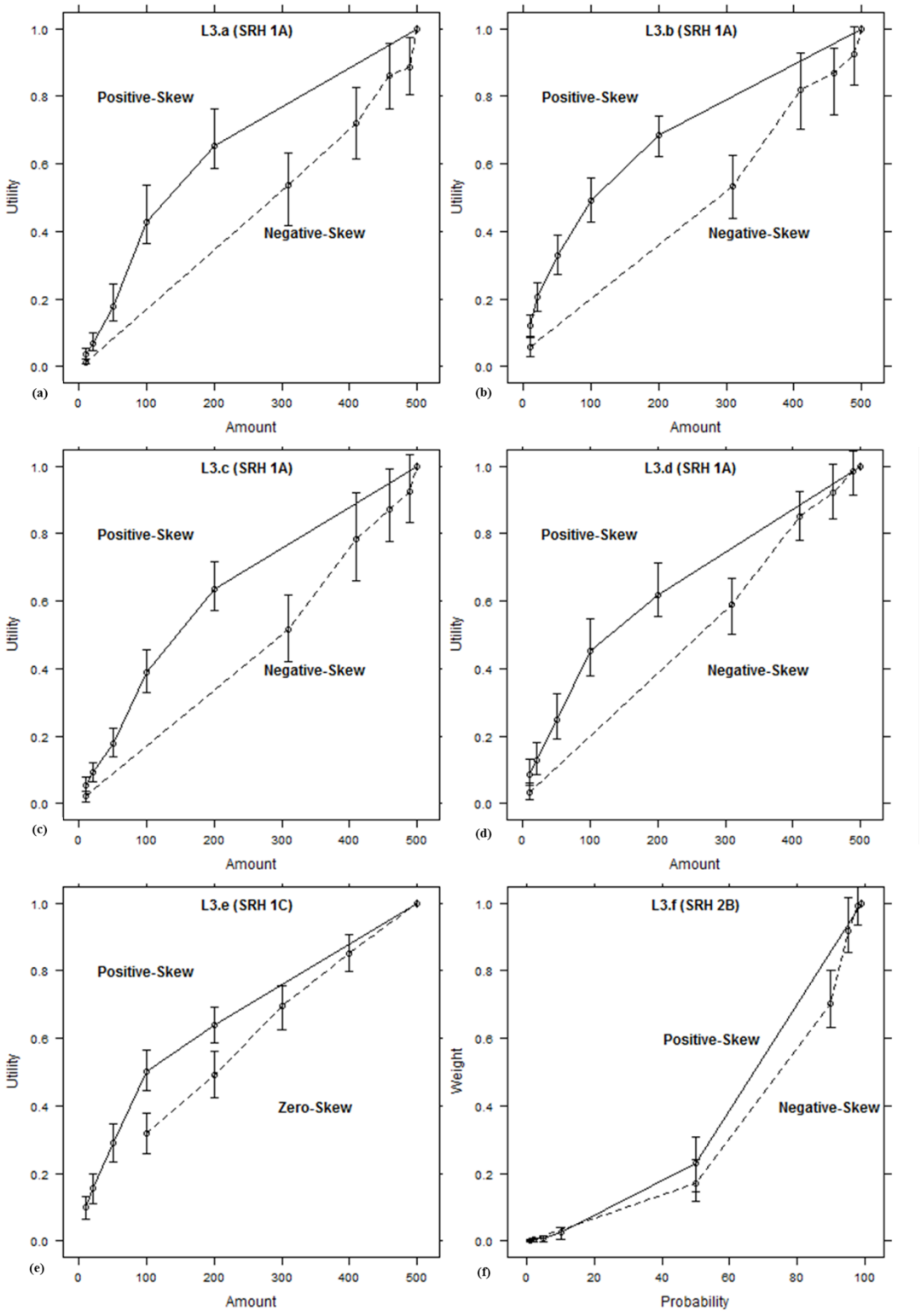

Figure B1. Revealed functions from the replications of SRH in Level 3 using two models instead of one model. Error bars are 95% confidence intervals.