Decision by Sampling and Rank Order Effects in Value Judgement and Decision Making

Timothy Leslie Mullett, BSc., MSc.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

Abstract

This thesis uses the Decision by Sampling model as a basis for examining effects of rank order encoding in value judgement and preferential choice. A range of experiments are reported, and these employ a variety of methodologies including behavioural paradigms, eye tracking and functional MRI. The results show that when there are a relatively small number of values used during an experiment, participants encode utility based upon the rank order of a potential outcome within these values. By introducing different decision contexts where the experienced values have a positive or negative skew, an individual's utility curve can be made concave and risk averse or convex and risk seeking. These different utility curves can be produced within the same individual and same task simply by providing a contextual cue for each trial. Two fMRI experiments demonstrate the neural systems underlying this phenomenon. The results show that all regions of the reward network encode reward as a function of the reward's rank order within the current context. No region of the brain was found to encode a reward's absolute financial value.

Other experiments investigated choice and valuation in more complex decision environments. It was found that when the number of experienced values is significantly larger than working memory capacity DbS is a relatively poor predictor of behaviour. The Weighted ADDitive rule proved to be more accurate throughout. However, in multi-attribute choice experiments where one attribute had a manipulated distribution, individuals use and weighting of the attribute value was determined by rank order rather than its numerical value. The specific characteristics of this were found to be incompatible with an exemplar based model of recall and binary comparison to specific items. It was instead found to be compatible with non-exemplar, fuzzy trace theories of decision making which are based upon estimates of the distribution. Eye tracking during multi-attribute choice additionally shows that participants begin to attend more to their preferred choice as they near the point at which they respond. However, they do not attend more to the attributes which they weight more highly in their choices, questioning the validity of previous eye-tracking findings.

Acknowledgements

Firstly, I wish to thank Richard Tunney for his support, feedback and constructive criticism through all my years at Nottingham. Hopefully during our collaboration and frequent discussions I have inched him closer to approving of cognitive neuroscience and other methods used during my PhD. Also, I thank Peter Chapman for his insights when discussing issues relating to eye -tracking and for his patience when training me to use the equipment.

I thank Lauren Marsh whose support has helped get me through the toughest times of the last few years. Your confidence in my abilities was consistently greater than my own and convinced me to keep going. Also, to the friends and family who offered support and advice, whether it was by sharing struggles and strife's from their own PhDs or offering escape from statistics, data and most importantly from MatLab.

Most of all I owe thanks to my parents who have supported me through all my studies and endeavours in life, even when tackling far greater adversity in their own lives. It saddens me greatly that my Father cannot be here to share in the happiness and acheivements that my studies have brought me, because it is only thanks to him and my Mother that I have been able to achieve so much.

Contents

1. Chapter 1 – Literature Review

1.1. Introduction to Decision by Sampling: The Model and Its Rationale

A large number of descriptive models of human value judgement and decision making have been proposed and tested in recent decades, showing a shift away from the economic prescriptive models and towards a more explanatory psychological model. Many of these models have been shown to predict choices with impressive accuracy. Although these descriptive models can predict choice, this is only the end result of the underlying process. These models are often incapable of describing the stages or mechanisms that lead up to the choice being made and few make predictions regarding other properties of decision making such as stochasticity, reaction times or priming phenomenon. A complete model of choice and valuation must be able to explain the process as a whole and in recent years there has been significant progress towards such a model.

The model which will be the main focus of this thesis is Decision by Sampling, hereon DbS (Stewart, Chater, & Brown, 2006). What makes the model so appealing is that DbS relies solely upon basic psychological processes. The complexity of its predictions is not a result of the process, but of the environment in which decisions are made and the distributions of values an individual has previously experienced. DbS predicts that when assessing the utility of a value, a random sample of previously experienced values is drawn from memory. Each of these values is compared to the value being considered. If the item under consideration is better than the one drawn from memory then the item's score is incremented by 1. If the value is no better than the sampled item then its score is not incremented. This essentially calculates the value's rank order within those previously experienced by the individual.

Consider a simplified example: An individual is shown or offered a potential gain of £5 and a potential gain of £10. To calculate the utility of these potential gains the individual then recalls a sample of previously experienced values from memory. Let us assume that the individual samples £1, £3, £4, £7, £15 and £30 and compares these to each possible outcome. The DbS score of £5 is 4, because there are 4 sampled items with a lower value. The score for £10 is 6, so according to DbS its utility is only 50% greater, despite the 100% increase in financial value. The DbS score can be plotted against value for each item in the sample to reveal the predicted utility curve (Figure 1.1). Although the simplistically small sample means the curve is not smooth, it is clear that the function produced a concave shape similar to that of Expected Utility Theory (von Neumann & Morgenstern, 1944). This is the result of the skewed distribution of the sample. The greater frequency – or over-representation – of smaller values means the relative rank rises quicker than in the upper section of the value range, where high values are comparatively rare. As we shall see, this sensitivity to skewed distributions is a characteristic fundamental to the accuracy and success of DbS.

For the vast majority of choices and comparisons, each option has more than one quality or attribute. In a simple financial gamble, these are payout value and probability. DbS predicts the same mechanism occurs for each attribute. Values are sampled randomly from a single attribute scale and then compared with the attribute value of the item under consideration. So when considering a gamble, a payout may initially be recalled and if its value is smaller than that of the current item then the DbS score will be incremented by 1. Then a probability may be recalled from memory and this will be compared to the probability of the current item. If the current item is favourable then its DbS score will again be incremented. Attribute values will continue to be sampled randomly from either probability or payout. Note also that there is only one DbS accumulator for each item, meaning that the effect of each attribute is additive and equal.

Let us consider a choice between £10 with a 0.5 chance otherwise nothing or a certain gain of £5. Assume that the same payout values are sampled as above, but that the following probabilities are also recalled: .0, .1, .2, .3, .7, .8, .9 and 1. For the risky option, the DbS score of the payout (6) is added to that of the probability (4) to equal 10. The same is performed for the certain option (4 + 8) for a total DbS score of 12. Therefore DbS predicts that the safe option would be preferred. Note that the sample of probabilities recalled can also have a significant effect upon DbS's weighting of chance.

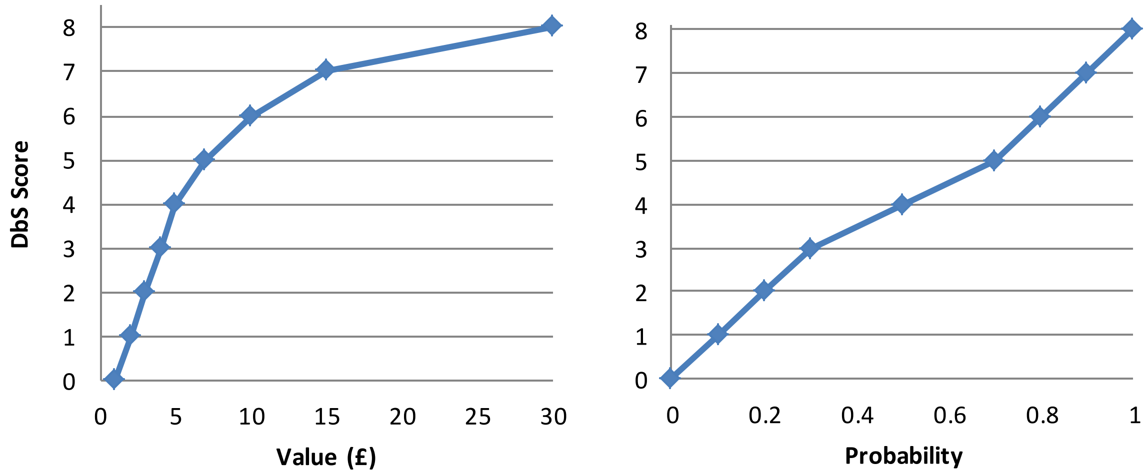

Figure 1.1 The DbS scores for hypothetical samples of values (left) and probabilities (right), plotted to reveal predicted utility and weighting functions.

The concave utility function shown in Figure 1.1 is the result of a significant positive skew in the sample used to create it. If the sample were equally represented at all points along the scale then utility would be a linear function and if the skew were negative then the curve would be convex. Therefore, it is a critical prediction of DbS that if individuals exhibit a concave utility there must be a significant skew in the distribution of values an individual experiences. Specifically, individuals should experience small financial transactions more frequently in daily life. Stewart et al (2006) obtained current account information from a major UK high street bank. By plotting the total number of credit transactions of each value on a log scale it is clear that individuals experience a disproportionate number of small gains (Figure 1.2). Therefore this real world pattern means DbS predicts the risk averse behaviour exhibited by the majority of individuals, without relying on an opaque mathematical weighting function (Figure 1.3).

Another pattern is evident when the debits are plotted in the same way as for gains: The skew is much more extreme for debits than credits (Figure 1.2). This is because people are more likely to experience larger regular credits such as a monthly/weekly salary and smaller regular losses such as grocery shopping or petrol purchases. This has the important consequence that losses are discounted more steeply than gains, so DbS also predicts loss aversion. This is something which is not possible without assuming different utility functions for gains and losses (Rabin, 2000) and also means that DbS can reproduce the utility curves central to so much of the success of Prospect Theory (Kahneman & Tversky, 1979).

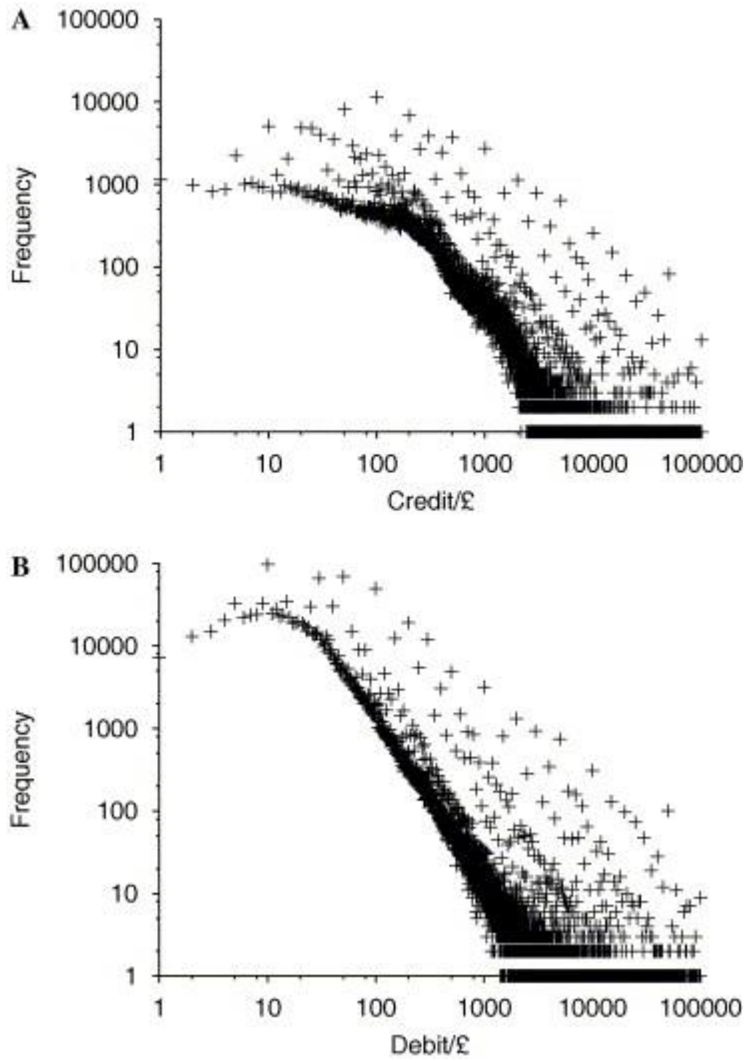

Figure 1.2 The frequency with which specific values are credited and debited from individuals' current accounts. Both show a significant positive skew, leading to risk aversion in DbS. The skew is also more severe for debits, meaning DbS also predicts loss aversion. Reproduced from Stewart et al. (2006).

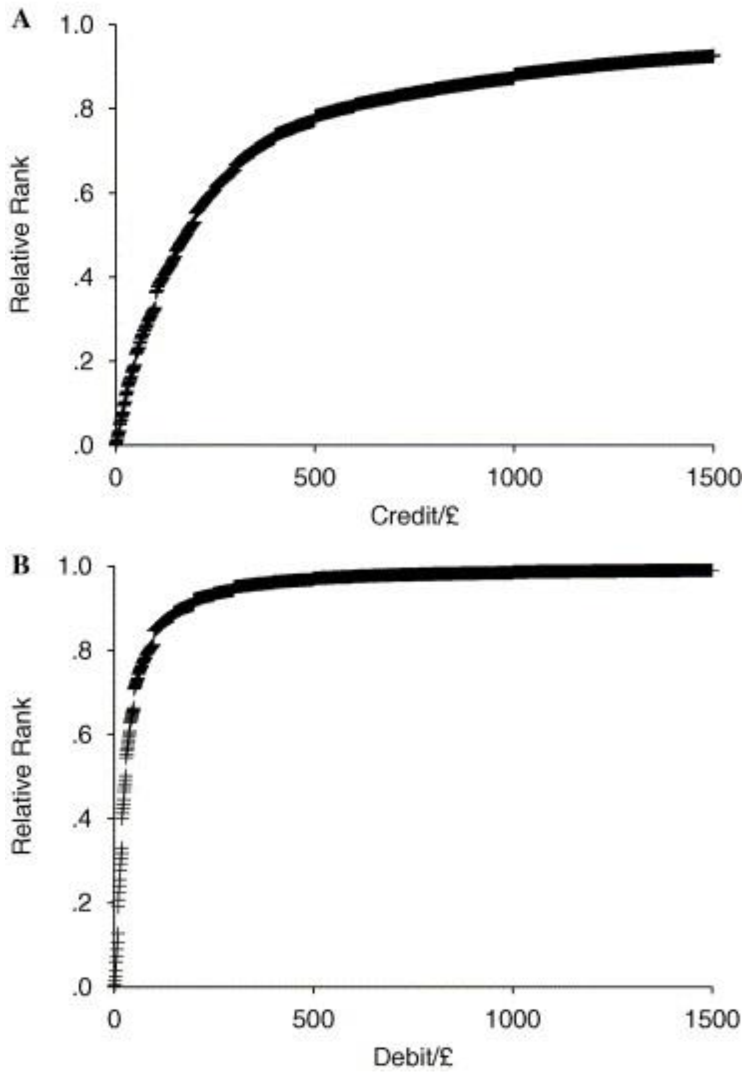

Figure 1.3 DbS's predicted utility curves as calculated using current account credits and debits. Reproduced from Stewart et al. (2006)

It has been demonstrated that there is no shape of utility curve which can explain or predict characteristics of choice – especially risk aversion – without also assuming a curvilinear weighting of probability (Abdellaoui, 2000; Abdellaoui, Barrios, & Wakker, 2007). If DbS is to fully account for prospect theory and match its impressive predictive accuracy for choices between financial gambles then the model must be able to predict the overweighting of small probabilities and the underweighting of large ones. DbS will show this pattern in an environment where very high and very low probabilities are experienced more often than probabilities in between. To also capture the risk averse pattern of underweighting a 50% chance, the distribution also needs a slight bias towards higher probabilities.

Unfortunately there is no real world analogue for the experience of probabilities which is as good a measure as bank transfers are for experienced gains and losses. However, there are still many measures which give valuable insights into individuals' experience and perception of risk. One of these is the frequency with which different chance or probability related words such as "likely" or "doubtful" are used in natural language. There is a significant existing literature which measures the perceived numerical equivalent of these words (Budescu & Wallsten, 1995). The numerical values can then be used to calculate the rank order of these words. In order to create the weighting function predicted by DbS, their frequency of use in natural language can then be used to calculate their relative rank and a weighting function can be plotted in the same manner as for the utility curves shown above (Figure 1.4). This shows high and low probabilities are indeed under and over-weighted respectively. There is also a slight bias towards probabilities greater than 50%, meaning that the function predicts risk aversion and crosses the diagonal at $p<0.5$.

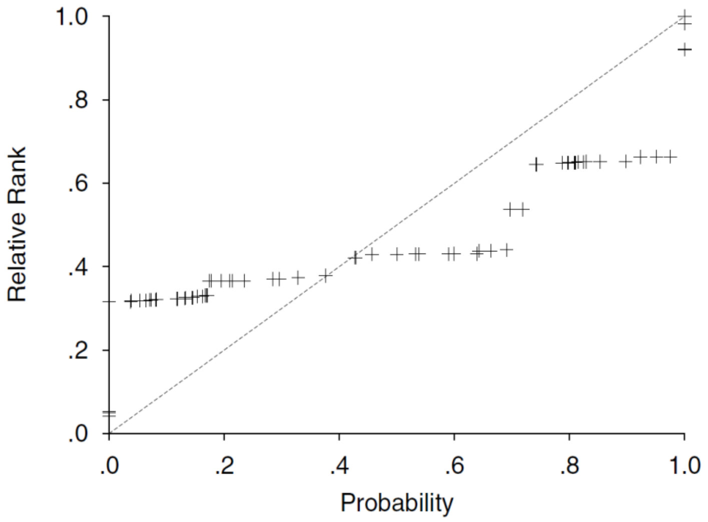

Figure 1.4 DbS's predicted probability weighting function as extracted from relative frequency of risk related descriptions in natural language. Reproduced from Stewart et al. (2006)

There are also other real world patterns which support DbS's use of prior experiences. The most significant is the over-reporting of rare events and of very likely events which failed to happen: Although lottery winners are regularly photographed in newspapers, those who bought a ticket and did not win are inevitably overlooked. Hence, individuals overestimate their chances of winning the lottery. The fear of terrorist attack has been found to be higher among those used to living in the relatively safety of the USA, than it is among those living in Israel (Yechiam, Barron, & Erev, 2005). This is despite the residents of Israel having far more direct, personal experiences of such events. This mismatch can be attributed to the fact that such attacks happen so often in Israel that individual events receive little or no attention in the news media. Whereas they are so novel and rare in America that any occurrence - or even a suggestion that an attack may occur - is investigated, debated and reported for protracted periods. What adds to the effect both in cases of lottery winners and terrorists is that the rare events are not only over-reported, but more salient. This means that they are more likely to be remembered and recalled when an individual is sampling prior experiences (Brown & Matthews, 2011; Brown, Neath, & Chater, 2007). But see Pachur, Hertwig, and Steinmann (2012), for evidence that reports of others experience have minimal effects upon judgements.

1.2. Experimental Evidence

1.2.1. Memory Constraints and Sample Size

There is now a growing literature, much of it from research on heuristics, that suggests the characteristics of human memory actually aid accurate decision making (Schooler & Hertwig, 2005). Many of these characteristics are assumed to be limitations because they preclude perfect recall. However, in many situations the patterns of experience in the real world interact with the “limitations” of memory to result in better recall and sampling of events and information which are more important to adaptive decision making (Anderson & Schooler, 1991). As will be covered below, there are now several studies which support the hypothesis that imperfect memory and recall results in more accurate judgements. However, these are still contentious and there are equally convincing findings suggesting that memory constraints lead to sub-optimal behaviour. Importantly for DbS, there is one common argument from both sides of this debate: that memory *does* impact decisions.

The suggestion that such complex decisions are made using simple comparisons with a potentially small sample of alternatives may seem surprising. This is particularly so when one considers the degree of discriminability humans are capable of and the confidence introspection often attaches to decisions. However, despite individuals often having high confidence in their judgements, there is significant stochastic noise in choice and individuals regularly make different choices when the same dilemma is presented multiple times (Glöckner & Pachur, 2012; Hey, 2001; Loomis, 1990; Mosteller & Nogee, 1951). DbS with a small sample size explains this effect simply as different items being retrieved from memory for each decision. When values experienced in a task are drawn from idealised distributions, it is more likely that a small random sample will accurately represent the population distribution. Experimental evidence shows that individuals' decisions and judgements become significantly more accurate as a result (Giguere & Love, 2013). This pattern is not predicted by larger samples or by a fuzzy trace account where estimated meta-information is stored and updated (Brainerd & Reyna, 1990; Kühberger, 1998).

A review by Juslin, Winman, and Hansson (2007) examined the findings of a large number of previous findings in the judgement and decision making literature. They concluded that in situations where there were sufficient previous experiences, results were compatible with a strategy of drawing small stochastic samples and then behaving as though these were accurate representations of the true distribution of the environment. Other studies also demonstrate that individuals made decisions based upon the observed distribution of values (Pachur et al., 2012) and that this is true even when participants are aware that these observations are not true representations of the underlying distributions (Feiler, Tong, & Larrick, 2013). There is also evidence that when in a novel environment, individuals at first rely upon a simple mean of the values or information they have experienced. But once a sufficient number of items and values have been observed they quickly switch to a strategy of drawing small stochastic samples from previous experiences (Lindskog, Winman, & Juslin, 2013).

Other research has explicitly investigated the link between short term memory and sample sizes. It has been shown that in open ended decisions from experiences tasks individuals with larger STMs sample more potential outcomes (i.e. collect more information from the environment), base their responses on relative frequencies and make more accurate judgements (Rakow, Newell, & Zougkou, 2010). Those with smaller STMs are found to rely on the same qualitative strategy, but use smaller samples, both in explicit information search and covert sampling from memory. Other tasks have used forced sampling and then inferred participants' use of information and sampling from their responses. These find that individuals with larger STM capacity sample more information and make more accurate probability estimates by virtue of considering more counterfactual events (Dougherty & Hunter, 2003a). Increasing working memory load during encoding and response in a relative frequency or a probability estimation task also increases errors in subadditivity in the manner predicted by a sampling model. The memory interference causes individuals with high capacity STM to behave more like those with low capacity (Dougherty & Hunter, 2003b). This effect of working memory is also shown to be significantly larger for judgements of probability of mutually exclusive events occurring than for pure frequency estimates (Sprenger & Dougherty, 2006), suggesting that the latter requires less information sampling and is less cognitively demanding.

Despite the evidence suggesting that the amount of information sampled is constrained by WM, this is not explicitly predicted by DbS. The DbS model predicts that only accumulator values need to be held in WM and that items can be sampled serially one at a time from memory until sufficient evidence has been accumulated to make a decision. However, as DbS does not posit an explicit stopping rule it is equally plausible that samples are drawn until WM is full and then the decision is made based upon this sample. The actual sample size and the stopping rule are factors to be investigated in the later chapters.

A mathematical consequence of relying upon small samples is that it significantly underestimates variance. When using the variance of a sample to estimate that of the population it is necessary to perform a correction (N/1-N). If individuals rely upon small samples then this correction must be applied in some way, otherwise individuals will systematically underestimate the variance within a distribution. Evidence shows that individuals do reliably underestimate the variance of an underlying distribution so cannot be applying such a correction (Hertwig, Barron, Weber, & Erev, 2006; Kareev, 2003). Furthermore, the effect is larger in individuals with smaller capacity STM demonstrating the increased bias resulting from smaller sample sizes (Kareev, Lieberman, & Lev, 1997).

Another consequence of small samples is a counter intuitive increase in sensitivity when detecting a correlation between two scales. Although this sensitivity is generally only representative of a type 2 error, it still provides a useful diagnostic tool in examining how individuals make their decisions. In one study, participants were shown a series of values from two different sources and then asked whether they believed there was any relationship, or correlation, between the two. Participants were then split into those with high and low short term memory capacity and it was found that the latter were more likely to report a correlation (Kareev et al., 1997). It should be noted that this also suggests individuals with a smaller STM will be more sensitive to change in the environment and a change in correlation. However, this pattern is not found (Gaissmaier, Schooler, & Rieskamp, 2006). The authors suggest that rather than STM constraints changing the sample used by individuals, it may instead promote different and simpler strategies or heuristics. An alternative account is that there is little bias towards sampling more recent items. Therefore, although individuals with smaller STM's are more sensitive to correlation, the equal representation of older experiences in the samples means that they would be no more sensitive to a change over time.

The hypothesis described above is questioned by studies of decisions from experience which do show a significantly larger effect for the most recent experiences. This is unsurprising given recency's importance within memory research (Anderson & Schooler, 1991; Ebbinghaus, 1913; Malmberg & Annis, 2012). Although effects of preceding stimuli are recognised in decision making research (Braida et al., 1984; Hogarth & Einhorn, 1992; Matthews & Stewart, 2009; Mori & Ward, 1995; Tversky & Kahneman, 1974), they are not predicted or explained by standard models such as Cumulative Prospect Theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). The fact that DbS is built entirely upon memory recall means that one would expect a significant bias towards sampling of the most recent values and items. Therefore the model does predict a recency effect and accounts for the anchor and adjust phenomenon. A series of experiments demonstrated that providing individuals with incidental values shortly before they answered a related dilemma manipulated choices for value, probability and delay discounting (Ungemach, Stewart, & Reimers, 2011). What was particularly interesting was that the results for value could not be explained by anchor and adjustment. They were best described by the relative rank of the values in the choice alternatives compared to the prior incidental values. Though see Matthews (2012) for a failure to replicate the effects in delay discounting.

Order effects have also been found in choices between serially experienced alternatives for non-financial decisions. When five identical glasses of wine were sampled sequentially, individuals had a distinct preference for the first and last options (Mantonakis, Rodero, Lesschaeve, & Hastie, 2009). When the second wine is sampled and compared to the first an individual may say they prefer the first. When the third wine is sampled, the most salient memory will be that the first wine has already compared favourably to one alternative, meaning the first wine's score is high according to DbS. As an individual continues tasting wines this score will be incremented and it becomes increasingly unlikely that a new wine will be preferred. However, if the individual does not prefer the first wine to the second, then it is less likely that a runaway favourite will accumulate a score in this manner. Therefore when the last wine is sampled, it is compared primarily to the one immediately preceding it as well as the fact that none of the previous experiences have a large DbS score. Therefore the chance of preferring the final wine is greatly increased. Furthermore, when two pleasant or high value items are considered one after the other then the second is generally preferred. In this case the memory of the preceding item experiences a regression to the mean. Hence, the effect reverses when two undesirable options are presented (Biswas, Grewal, & Roggeveen, 2010).

Many of these results come from the literature on decisions from experience but DbS is primarily a model of decisions from description. DbS essentially combines explanations from these two modalities. Decisions from experience are the judgement of a single choice option or event based on previous experience of that specific choice option or event. One interpretation of DbS is that the same processes which apply to judging an event, can also be generalised to judging a description. So instead of sampling occurrences of option A, an individual samples occurrences of probabilities or of financial gains. This is arguably a more ecologically valid description as judgements from description have only become common relatively recently in human evolutionary history and descriptions of probabilities even more so. Hence, evidence that individuals are more accurate at making decisions when descriptions are based upon relative frequencies compared to descriptions using probabilities (Cosmides & Tooby, 1994; Sprenger & Dougherty, 2006).

1.2.2. Effect of Available Alternatives

There are a large number of phenomena where it is not previous experiences that modify individuals' decisions and judgements but the set of current alternatives. One of the simplest of these is the dominance effect (Ariely & Wallsten, 1995; Huber, Payne, & Puto, 1982). Imagine a situation where an item can differ on two attributes, X and Y. An individual has to choose between items A and B. A has a high value on X, but a low value on Y, whereas B has a high value on Y and low value on X (Figure 1.5). The relative differences on scale X and Y have roughly equal importance to decision makers, meaning that individuals are split equally between the options. However, item C is then introduced to the choice set and is slightly worse than item A on both X and Y. A dominates C because it is better in all possible ways. Now when individuals are asked to choose, a significant majority prefer option A and the proportion of choices for option B decreases significantly, despite the relative difference between the two staying the same.

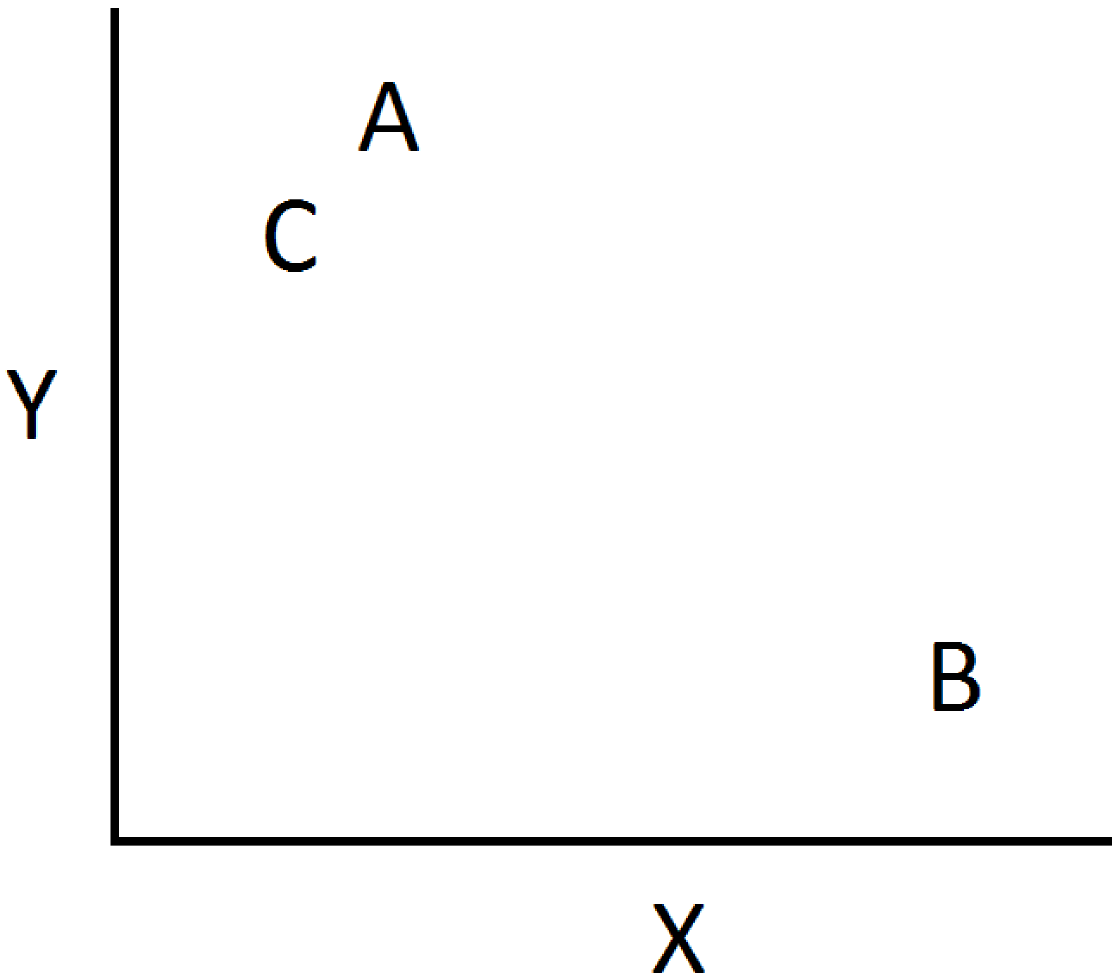

Figure 1.5 The dominance effect. A,B&C are multi attribute items, varying on scales X and Y.

DbS explains this as the simple result of binary comparisons between the available options. When sampling alternatives the most salient items will be those presented as direct alternatives as these require no memory retrieval and minimal cognitive effort. Therefore option C has a DbS score of only 1 because it is better than one other option on one attribute. Option A has a score of 3 because it is better than C on both attributes and better than B on one. Option B has a score of 2 because it is better than both A and C on one attribute, but is worse than both on the other. Therefore, option A is judged most favourably and chosen most frequently. This effect also has longer term consequences. When a previously dominated item such as C is later seen outside of this context it is valued lower than when a dominating option such as A is seen (Biswas et al., 2010). This suggests that the items an option was originally seen with are relatively more likely to be sampled when the option is seen again. A phenomenon supported by memory research (Godden & Baddeley, 1980; Tulving & Thomson, 1973; Tunney, Mullett, Gardner, & Moross, 2012).

The distribution of items in a decision environment also shows more complex and nuanced effects upon choices. Judgements of the severity of road accidents is significantly modified by the distribution of other accidents presented at the same time (Robinson, Loomes, & Jones-Lee, 2001). If the context within which valuations are made is disproportionately comprised of high severity accidents then items in the centre of the scale received lower severity estimates. This is true even when the most and least severe items are present in all conditions, meaning that the overall range is constant. The effect is also found when judging raffle tickets. Confidence of winning was not simply dependent upon what proportion of the tickets were held by the individual, but also upon how the others were distributed (Windschitl & Wells, 1998). If they were split between a large number of individuals then participants were more confident as their chances compared well to many other people. However, if all the tickets were held by one other person then participants were less confident of winning as the only other person with which they could compare themselves had a greater chance of winning. Showing additional individuals with very low probabilities of winning served to increase individuals' confidence of winning themselves, despite reducing their actual chance of doing so (Windschitl & Chambers, 2004). These effects also interact with STM capacity. Adding more low probability alternatives has a greater effect upon judgements in individuals with a larger STM capacity (Sprenger & Dougherty, 2006). However, it should be noted that some argue these effects are not explained by rank order, instead suggesting an anchor and adjust effect with the next most likely outcome (Windschitl & Young, 2001).

It is not just the distribution of the values themselves that can affect decisions. The distribution of possible response options has a similar effect (Stewart, 2009). The average value estimate for a gamble or item can first be elicited using an unbiased method. Individuals are then asked to choose the value of an item from a set of possible responses. If the possible responses start at a value slightly below the average value and extend very far above it then participants' responses are higher. If the highest possible response is only slightly above and the lowest is far below, then responses are significantly lower (Stewart, Brown, & Chater, 2005; Stewart, Chater, Stott, & Reimers, 2003). The effect is also apparent when the range of options is kept constant by using the same maximum and minimum values. The values between can be positively or negatively skewed, such that low or high values are over-presented. Estimates are subsequently higher and lower, in the direction predicted by DbS and rank encoding (Birnbaum, 1992).

The effects of possible response options have significant and possibly damaging real world effects. One decision making task where this has been explicitly examined is allocation of pension funds. When saving for a private pension, individuals have to decide what proportion of their funds they wish to assign to different bonds, cash and stock options. The most common strategy is to split their funds equally between all options available in their particular pension scheme. However, individuals do not adapt this strategy depending upon the split of different options (Benartzi & Thaler, 2001; Vlaev, Chater, & Stewart, 2007). So if a particular pension scheme has many stock options, but only one cash option, then the individual will use the decision environment as a cue, spreading their assets "equally", but then leaving themselves exposed to a disproportionately high degree of risk.

1.2.3. Rank Ordering as a General Strategy

Evidence of judgements by rank order comes from many sources and domains, not just financial decisions or the JDM literature. The earliest evidence for rank order-encoding came from psychophysics, primarily volume judgements. Although human hearing allows for very accurate discrimination between the volume of two sounds or the frequencies of tones, accuracy is poor when individuals are asked to provide an estimate of the absolute magnitude of a single example (Garner, 1954). These early findings inspired a line of research investigating why absolute judgements were so unreliable and this led to theories describing how these judgements were made. The most successful of these has been Range-Frequency Theory (hereafter RFT; Parducci, 1965).

RFT posits that when an individual assesses an item with an absolute magnitude or value and assigns it to one out of a set of ordinal categories they use both the range of values and the frequency with which items are assigned to each category. For example, when judging the loudness of a series of tones using the labels, very quiet, quiet, medium, loud and very loud. If tones range between 0dB and 100dB then a simple range adaptation account will split the scale into 5 bins with a width of 20dB each. However, if the stimuli set contains a large number of quiet tones then the frequency component of the model will adjust in order to keep relatively equal numbers of stimuli assigned to each category. Therefore, a smaller range of the scale will be assigned to the categories "very quiet" and "quiet" meaning they cover less of the scale. The range covered by the "medium" category may move lower down the scale, while "loud" and "very loud" essentially grow to encompass more of the scale. The overall effect is that tones in the mid-range are assigned to louder categories than they would be in a non-skewed environment. The obvious mechanism for extending the model from categorical judgements to absolute judgements is a model of valuation by rank order, with the range of possible responses setting the maximum and minimum of the scale. This maintains the fundamental properties of the range-frequency account, whilst also allowing it to be applied to judgements such as certainty equivalence. It also means that the model is very similar to the transform applied to value in DbS when an individual is making a decision rather than a judgement.

The early findings which inspired RFT have now been extended and many of them support absolute judgement by comparative mechanisms (Stewart et al., 2005). The rank of a tone's volume or loudness amongst preceding items has a significant effect upon the judgement of subsequent stimuli (Stewart & Brown, 2004). The effect is also found in the judgements of many other simple perceptual stimuli (Stewart, Brown, & Chater, 2002). These effects have also been shown in domains which blur the lines between psychophysics and cognition such as subjective ratings of pain (Watkinson, Wood, Lloyd, & Brown, 2013). It has even been shown in enjoyment and perception of music, with RFT predicting judgements' of the most accurate and most pleasing tempo at which to listen to Beatles songs (Rashotte & Wedell, 2012).

Effects of frequency and rank have been shown in a large number of domains and judgements. Many of the documented examples are in social judgements about the self. For example, individuals' judgements of their own happiness or depression is explained by comparisons with other known individuals and an individual's estimated rank within this sample (Melrose, Brown, & Wood, 2013). This is also true of judgements about different personality traits (Wood, Brown, Maltby, & Watkinson, 2012) and a number of other social scales (Galesic, Olsson, & Rieskamp, 2012).

Individuals' estimates of how their earnings compare with those of the general population are best explained by their rank within their peers and social group (Brown & Matthews, 2011). Employees well-being has also been found to rely upon how their earnings compare with those around them, rather than the financial amount (Brown, Gardner, Oswald, & Qian, 2008) and this is also true of general satisfaction with earnings (Boyce, Brown, & Moore, 2010). Particularly intriguing are findings regarding mental health issues and the lower psychological well-being associated with low socio-economic status. These risks are better predicted by an individual's socio-economic rank within their local community than it is by their actual wealth and resources (Wood, Boyce, Moore, & Brown, 2012). This suggests that the risk factor associated low SES is not the result of lack of resources, such as access to good nutrition, exercise etc. but is at least in part a psychological effect driven by comparison with others.

Comparative and rank effects also predict judgements of potentially harmful behaviours. Individuals' judgement of their actual alcohol consumption and the estimated harm of their own drinking is explained by their rank within other drinkers in their social circles (Wood, Brown, & Maltby, 2012). This is also the case for behaviours which have a positive effect upon health, such as amount of exercise and the predicted positive impact of their exercise (Maltby, Wood, Vlaev, Taylor, & Brown, 2012). Even offers of help are judged in this way, with individuals displaying more or less gratitude depending upon how well an offer compares to other experienced offers (Wood, Brown, & Maltby, 2011). These findings have significant implications for promoting healthy behaviours and designing interventions. Whether such interventions or the design of decision environments can also aid better financial decision making, particularly in areas such as pension provision, is still a focus of much debate (Sunstein & Thaler, 2008).

1.3. Criticisms of Decision by Sampling and Open Questions

Despite the strong and growing body of evidence in favour of DbS's core properties and predictions, there are still valid criticisms which can be levelled against it. There are also a number of crucial questions and characteristics of the model which are currently open or unspecified. For DbS to become widely accepted as a model of human decision making and value judgement then these issues must first be addressed.

Perhaps the most surprising issue given the otherwise strong base of evidence is that there is currently no published test of DbS's predictive accuracy in financial decision making. Although individual phenomena have been closely examined and robustly demonstrated, there has not been a single study where DbS was explicitly modelled and used to predict behaviour. There are inevitable problems with such a test: the fact that DbS is a memory and experienced based model means that one cannot measure or control for experiences prior to beginning the experiment. Not being able to measure these and enter them into the model will inevitably add noise. However, recall shows a steep forgetting curve (Ebbinghaus, 1913), effects of categorisation mean that items within the current task or block are most likely to be sampled (Brown et al., 2007) and the number of items sampled is potentially very small (Cowan, 2001). Therefore, the likelihood that individuals sample more temporally distant experiences from outside of the experiment is also likely to be very small. Given the inherent noise in human judgement, the effect of prior experiences is likely to have relatively little impact upon modelling accuracy.

Another issue which makes modelling DbS difficult is that it is underspecified. The model makes no predictions regarding a stopping rule or the size of the sample used to make decisions. STM capacity has been shown to be very small (Cowan, 2001) and numerous studies detailed earlier in this chapter suggest that decisions are modified by the size of STM. However, as DbS is an accumulator model it does not predict that all sampled values are necessarily held in STM at once. Instead values are sampled sequentially, allowing for a much larger number of comparisons. There is already evidence that the differences in decision making that correlate with STM may not be a direct result of sampling or parallel sample representation (Gaissmaier et al., 2006). Therefore any modelling of DbS would have to estimate the size of the sample with which to make predictions.

In addition to not specifying the size of the sample, DbS fails to specify a neurologically and psychologically plausible mechanism by which previous experiences are recalled and sampled. The assumption stated in the original model specification is that all prior experiences have equal random chance of being recalled (Stewart et al., 2006). This seems to be because the majority of the evidence presented in the paper relies upon long term measures of experiences and events. However, more recent and more targeted experiments have demonstrated the importance of recency, saliency and similarity. Therefore an unavoidable question is how memory phenomena and the differential likelihood of experiences being recalled are integrated into the DbS model.

One characteristic of the model which is fully specified is how items with multiple attributes are judged and how the different scales are integrated. Despite this, little attention has been paid to multi-attribute valuation, with evidence instead coming from single attribute tasks or financial gambles. This is an issue which can be easily addressed by modelling DbS's predictions. The model currently predicts that information is sampled randomly from each attribute with equal chance. However, it seems that in most decisions individuals do not weight information equally (Bröder, 2002; Mellers, 1980; Pitz, Heerboth, & Sachs, 1980; Westenberg & Koele, 1994), although the impact of such weighting is debated (Dawes, 1979). It seems unlikely that these issues of information weighting will be overcome by using rank order encoding. Therefore, weighting parameters may need to be incorporated into the model. The most likely and parsimonious mechanism would appear to be preferential sampling of the attributes which are most salient and considered most relevant to the current decision.

1.4. Comparisons with Alternative Models and Interpretations

1.4.1. Prospect Theory

The distribution of experienced values in the world means DbS predicts utility curves that closely match those of Prospect Theory. It would seem a natural assumption that CPT and DbS therefore make the same or similar choice predictions. However, they are subtly, but importantly different. The utility curves calculated using DbS were based upon population level data. When the gains and losses of a large number of individuals, from very rich to very poor are pooled then the familiar utility curves are indeed reproduced. However, when an individual is making decisions they can only sample from their own experiences meaning DbS predicts different individuals will have very different utility functions, as the very wealthy and the very poor will have experienced very different gains and losses. Furthermore, an individual will only sample a relatively small number of these items for a single decision, meaning that the utility function will not be accurately represented during each choice. However, it will become apparent when averaging over a large number of choices. Furthermore, CPT predicts that utility functions are stable over time, whereas DbS predicts an adaptation and inherent updating of the information used to make decisions, meaning utility functions are dynamic and adjust to the decision maker's environment.

1.4.2. Heuristics and the Adaptive Toolbox

The debate surrounding the relative merits of heuristic accounts of decision making and models with strong mathematical components is one of the most fundamental within contemporary JDM research. A full discussion is beyond the scope of this work, but many issues of the debate are relevant to DbS and the experiments reported here. The central argument in favour of a heuristics account of decision making is that the mathematical computations in mathematically complex models are incredibly cognitively demanding. So much so that, for many models individuals are incapable of performing the relevant mathematical transforms when explicitly asked to do so. Generally speaking, evolutionary processes do not result in any processes more complex or energy demanding than is necessary or possible. Therefore it is argued that the application of simple rules, which approximate the results of complex mathematical strategies, are more psychologically plausible (Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2002; Gigerenzer & Todd, 2000).

DbS is a particularly interesting case because it arguably sits astride the debate. The simple mechanism of recall and binary comparison is a more plausible cognitive process than complex discounting. It is arguably a model of bounded rationality as it considers cognitive limitations and yet produces the similar results to complex mathematical weighting functions. However, as DbS is a relatively general cognitive model rather than a specific rule, it still does not fit well within the adaptive toolbox (Bröder, 2003; Gigerenzer & Selten, 2002) alongside the matching (Dhami, 2003; Snook, Dhami, & Kavanagh, 2011), recognition (Goldstein & Gigerenzer, 1999) or priority heuristics (Brandstätter, Gigerenzer, & Hertwig, 2006). A potential result of this is that DbS may be able to fit findings compatible with simplistic, rule based heuristics, whilst also explaining the complex utility functions of mathematical models. For example, if attributes are differentially sampled due to characteristics such as salience or similarity, then the model may be able to explain patterns of behaviour where individuals make complex use of information and appear to weight different attributes mathematically. Conversely, depending upon the assumed stopping rule, the model can explain effects compatible with matching and priority heuristics by assuming that the most influential attributes are sampled first or disproportionately more. If the accumulator then reaches its stopping rule, further attributes will not be sampled. Furthermore the most common explanation for recognition and representativeness heuristics is already that of preferential recall from memory (Goldstein & Gigerenzer, 1999; Kahneman & Tversky, 1972; Tversky & Kahneman, 1974).

1.4.3. Drift Diffusion

Drift diffusion is a stochastic model (or family of models) of choice which like DbS predicts that evidence for each item is accumulated over time until a stopping criterion is reached (Busemeyer & Townsend, 1993; Ratcliff, 2001; Ratcliff & McKoon, 2007). Where this model differs from DbS is in how the evidence is accumulated. Drift diffusion assumes that evidence for an item is accumulated at a rate dependent upon the relative differences in value between the options under consideration. So in a choice between two very similar items, the rate of evidence accumulation will be slow and decision times long, whereas in a choice between two items with very different values, evidence is rapidly accumulated for the more valuable item and decisions are fast.

There is a strong base of evidence for drift diffusion models, with many studies finding impressive predictive accuracy. However, a potential criticism is that many papers rely upon a generalised modelling procedure, averaging across trials and not capturing the potential signal on each choice. In addition, both decisions and response times are used when estimating model parameters and when testing the model's predictive performance. This is indeed a good method of testing drift diffusion as a process model, but it also means that comparisons with competing models are not very informative, as most do not make specific predictions regarding reaction times. The modelling also estimates a varying number of free parameters depending upon the specific implementation of the model (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) and in some cases the model becomes very complex. Although this is best avoided in cognitive models, there is evidence that the parameters have real psychological and neurological analogues. Specific parameters in the model are modified by relevant changes in task constraints (Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010), by neurological damage and aging (Ratcliff, Thapar, & McKoon, 2006; Starns & Ratcliff, 2010), and even sleep deprivation (Ratcliff & Van Dongen, 2009).

Research on drift diffusion models is becoming more and more sophisticated over time and this work is also becoming more relevant to DbS as both posit a system of evidence accumulation. The major difference is in what evidence is being accumulated: binary comparisons or absolute differences. One line of research which is particularly interesting is that examining attention effects. By tracking decision makers' eye gaze, it is possible to examine which item and attribute is being looked at and therefore attended to at any particular time during deliberation and decision making. This has enabled modelling of drift diffusion by providing a measure of which item evidence is currently being accumulated for (Krajbich, Armel, & Rangel, 2010). However, there is no reason this methodology could not also be applied to DbS. This would make it possible to compare the predictions of the two models. When attention is directed to one item in the choice set, is evidence accumulation best described by the attribute value's rank order within previous experiences? Or is drift diffusion's assumption correct that the accumulation rate depends upon absolute difference in value? If the latter then the issue remains of how this absolute difference is calculated, and whether a complex utility function is applied to it.

1.5. Thesis Framework

This thesis examines Decision by Sampling as a model of valuation and choice. It also assesses the more general role that rank ordering plays in judgement and decision making. Chapter Two uses a multi-attribute valuation task to examine the predictive accuracy of DbS when it is explicitly modelled upon the stimuli. It also tests predictions of range frequency theory and rank order in value estimates using modified distributions of stimuli values. Chapter Three extends these findings, examining the predictions of DbS by modifying the distribution of values upon individual attribute scales. It also examines the effect of prior experience by comparing value estimates in tasks where participants have prior knowledge or the stimuli are completely novel. Memory phenomena are also introduced to the DbS model to assess the most likely sample size used by individuals and whether it is possible to improve the models predictive accuracy by incorporating memory effects.

Chapter Four extends the results of chapter two into multi-attribute choice, comparing the effect of binary comparison heuristics such as Dawes rule (Dawes, 1979) and the effect of absolute value differences between options. The paradigm is also used in combination with eye-tracking to provide an estimate of relative attention. This allows drift diffusion models to be simulated using a novel modelling method then compared to DbS and simpler behavioural models. Chapter five focusses on neuroscience findings and uses fMRI to dissociate the effects of absolute value and rank order in the neural encoding of value. The effect of context upon neural encoding is also examined as well as how context is defined in different neural systems. The final experiment takes the findings from the neuroimaging experiments and demonstrates an analogous behavioural effect outside of the scanner. This experiment shows a cross-modal replication; something which is all too rare between the neuroscience and psychology literatures.

2. Chapter 2

2.1. Introduction and Overview

This chapter details two experiments which assess Decision by Sampling's applicability to the valuation of single items with multiple attributes and the performance of the closely related range frequency theory (Parducci, 1965). These experiments allow for modelling of DbS and the exploration of several potential modifications or additions to the model which are inspired by the literature on memory phenomena. The design of both experiments also allows for examination of learning rates during multi-attribute valuation tasks. Previous research has suggested that participants are unable to use feedback and learn the effect of variables when the decision environment is difficult or there are more than two pieces of information having an effect upon value (Harvey & Fischer, 2005).

Experiment one required participants to estimate the rental value of a series of apartments based upon a number of pieces of information. Feedback was provided in the form of the correct rental value after each trial. This allowed an examination of participants' learning and accuracy rates over the course of the experiment. The distribution of rental values experienced was also skewed. This means that the predictions of DbS and range frequency theory deviate significantly and systematically from the true rental values provided as feedback in the task. Experiment Two provides a further test of range frequency by presenting a distribution of rental values where a large portion of the range is under represented.

2.2. Experiment 1

2.2.1. Introduction

Experiment One was a multi attribute value estimation task. Participants were shown the details of a series of apartments and asked to estimate the monthly rental price. Apartments were chosen due to the existing literature using such items to investigate information search and multi-attribute decision making (Payne, 1976). They are flexible and appropriate stimuli because they are gender neutral (unlike cars or shoes for example; Matthews & Stewart, 2009) and they have a large number of potential attributes which can be numerically quantified. Furthermore, our undergraduate participant pool would have reasonable understanding of the items as rental tenants but would not be experts with the strong, predefined expectations of estate agents or land-lords.

The most basic aim of the experiment was to assess participants' learning and to examine whether they could accurately extract weighting functions or information about individual attributes and thus improve the accuracy of their item valuations. There is debate surrounding individuals' ability (or inability) to learn the relationships between individual cues and overall value. Some studies show that individuals have serious difficulties learning from simple outcome feedback when there are more than two attributes (Olsson, Enkvist, & Juslin, 2006; Todd & Hammond, 1965). Others suggest that participants can only learn to make accurate use of cue-outcome relationships when provided with detailed feedback which often explicitly reveals the relationship between cue and output (Balzer, Doherty, & O'Connor, 1989). However, much of this evidence comes from multiple cue probability learning tasks where outcomes are binary and probabilistic. There are a number of studies which have used stable and deterministic environments with feedback given on a continuous scale. These show that participants can achieve high accuracy (Brehmer, 1994; DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004; Mellers, 1980). For a review see Harvey and Fischer (2005).

The stimuli in this experiment were stable and each attribute had a deterministic mathematical relationship with item value. In order to provide the most powerful test of learning possible, the same 100 items were presented twice and in the same order, meaning participants essentially repeated the experiment. Therefore the accuracy of value estimates at each presentation of the same item could be compared to provide an accurate measure of learning. It also means that once an asymptote is identified, it is also possible to assess the long-term stability of value estimates.

In addition to this the experiment allowed a direct test of both DbS and RFT. RFT does not make specific predictions regarding the interpretation and integration of attributes onto single internal psychological scale. However, it does make predictions about the transformation from this internal value to a financial judgement, i.e. the response value. The distribution of rental values used in the experiment had a significant positive skew. This results in RFT predictions being significantly and systematically different from the true rental value. DbS makes specific predictions about the interpretation of individual attribute scales and the methodology here allows for modelling of DbS. The predictions of DbS can then be compared to participant responses and the accuracy compared to that of simply using true rental value.

It is also possible to examine a number of potential improvements to the DbS model. The model currently predicts that all previous experiences have an equal chance of being sampled. However, previous results have shown that participants can adjust to a skewed environment after a relatively small number of trials and that the most recent experiences have a significant effect upon decisions (Stewart, 2009). Furthermore, findings in the recognition memory literature show that more recent experiences are more likely to be recalled from memory (Ebbinghaus, 1913). Others show both a recency and primacy effect (Glenberg et al., 1980; Mantonakis et al., 2009). Therefore the modifications of DbS tested here incorporate a weighting function to relatively over-sample either more recent items or both early and recent items. If either of these modifications shows an improvement then it would demonstrate that this short term adaptation is a fundamental aspect of a memory based decision making process. It would also rule out anchor and adjust (Tversky & Kahneman, 1974) as an explanation of previous findings.

2.2.2. Method

2.2.2.1. Participants

Participants were 12 Psychology undergraduates who were paid £5 for participation. Eight of these were female and four male, with an average age of 19. 8 (S.D. = 2.7).

2.2.2.2. Stimuli

The stimuli consisted of 100 apartments. Each apartment had five attributes: Number of Bedrooms, Number of Bathrooms, Floor Size, Land-Lord Rating and Distance from Town. Floor size was reported in square foot and distance from town centre in miles. Participants were informed that land-lord rating was a score between 0 and 10, calculated using feedback from previous tenants. Attribute values were created using the parameters detailed below. These were chosen so that items were a reasonably close approximation of the smaller apartments advertised to students in the Nottingham area.

- Beds – random integer between 1-4
- Baths – random integer between 1 and the number of bedrooms
- Square footage – base of 500, plus a random amount up to 500 (drawn from a boxcar distribution), plus an additional 50-250 for each bedroom (also drawn from a boxcar distribution)
- Land lord rating – continuous scale from 1 to 10, randomly drawn from a normal distribution with a mean of 7 and a S.D. of 3. Values drawn from outside of this range were re-sampled, to avoid a cluster of values at the extremes which would have occurred if values were instead rounded.

- Distance from town – values over 0.1m randomly drawn from a normal distribution with mean of 0 and S.D. of 2. Again, values less than 0 were re-sampled, resulting in a half-normal distribution.

The rental value was then calculated using a set of mathematical formulae. Essentially, the starting value of £220 was multiplied using the formulae below, so that each attribute had a specific effect upon value. The product of the first equation was then multiplied by the next and so on, so that the output had a weighted multiplicative relationship with the stimuli values. This explicit mathematical function means that there was an objectively correct value for each item and a deterministic relationship between attributes and rental values. Therefore it was possible to measure how quickly participants responded to feedback and to the decision environment in order to improve their accuracy. The weighting functions were piloted on a small number of students in order to test their plausibility and underwent a number of refinements before the experiment was conducted. In the experiment itself no participants reported that they thought the values were odd and anecdotally, several apparently believed that the data was drawn from real adverts.

Equation 1

$$1 + 6(beds - 1)$$

Equation 2

$$1 + \frac{3(baths - 1)}{20}$$

Equation 3

$$\frac{1.2}{0.5\left(\frac{750 + 150beds}{0.5sqft}\right)}$$

Equation 4

$$1 + \frac{(llord\ rating - 5)}{20}$$

Equation 5

$$1 + 0.3\left(\frac{2}{10dist}\right)$$

The parameters detailed above were used to create 90 stimuli. The remaining 10 had their attribute values specifically chosen such that there was an item with a rental value in every 10^{th} percentile of the total range. These items were then placed at the end of the series of 90 items and were used as critical items or probe trials. This is particularly informative because the 10 items could be used to test for any systematic differences across the range of values after being exposed to the skewed distribution of the items experienced immediately prior. The overall distribution of rental values had a range of £371 to £1638 and a mean of £801. The distribution also had a significant positive skew (Figure 2.6, skewness = 0.71). Thus if DbS is correct then the skewed sample of experienced values will result in systematic deviations from perfect calibration.

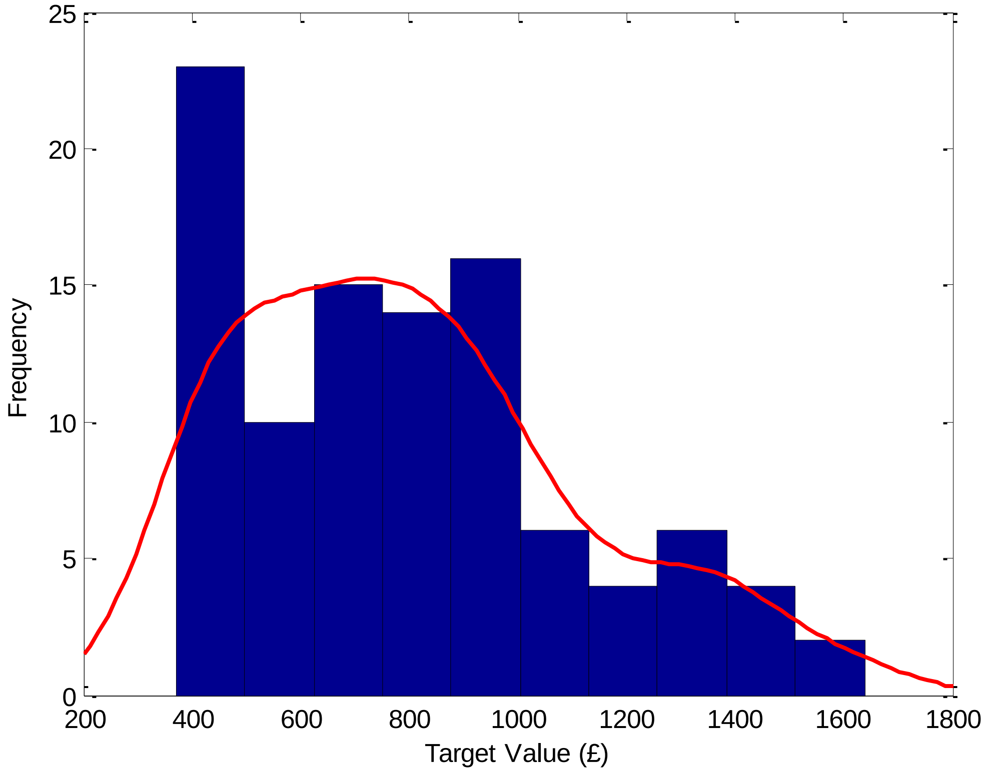

Figure 2.6 Distribution of stimuli values. Continuous estimates calculated using nonparametric kernel-smoothing with 100 samples.

2.2.2.3. Procedure

Participants completed 200 trials. This was composed of the 100 items described above. These 100 were repeated in the same order allowing analysis of the stability of participants' value estimates over time.

Participants were told they were completing a prototype estate agent training task and would have to estimate the potential rental values of a series of apartments based on a small amount of preliminary information. For each trial participants were presented with the details of a flat and responded with their value estimate. Once this had been provided, the true rental value was displayed for 2 seconds before the next trial began.

2.2.3. Results

2.2.3.1. Learning and Accuracy Rates

To assess accuracy over the duration of the experiment the percentage error rate was calculated for each trial. First the difference between each participant's estimate of value on each trial and the true rental value was found. Then the difference on each trial was divided by the true stimuli value in order to find the percentage error. This measure was used in preference to simple subtraction because a difference score would inflate the influence of more valuable stimuli. Figure 2.7 shows the change in accuracy over time and control for trial to trial variability a LOWESS smoothed curve was calculated and plotted in addition to the simple mean of the nearest 30 trials. Both were calculated using a window encompassing 15% of the data and the LOWESS regression used robust errors. This shows that accuracy is poorer early in the task but improves quickly, becoming relatively stable in the first half of the experiment. The percentage error then remains around 15%. To test for the stability of estimates over time, differences between value estimates for the first and second viewing of stimuli were used to calculate mean error percentage in the same manner (Figure 2.8). This reveals a very similar pattern but suggests that the difference in estimates stabilises slightly quicker: between the 30th and 40th trials.

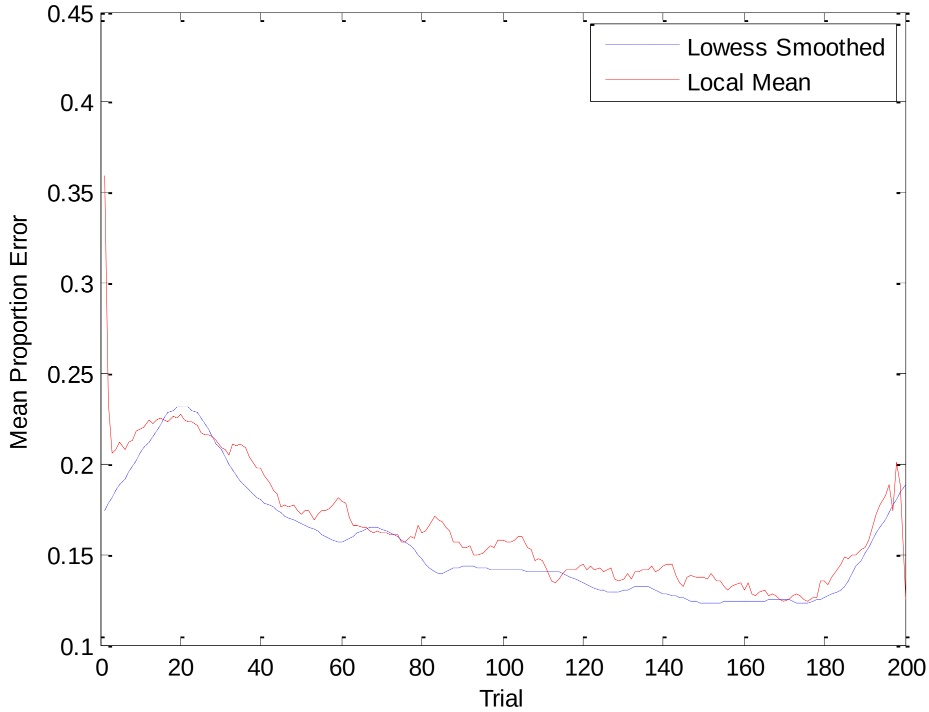

Figure 2.7 Mean size of the error as a percentage of the stimuli's target value over the duration of the experiment. Different lines show two methods of smoothing, a simple local mean and a LOWESS smoothing method.

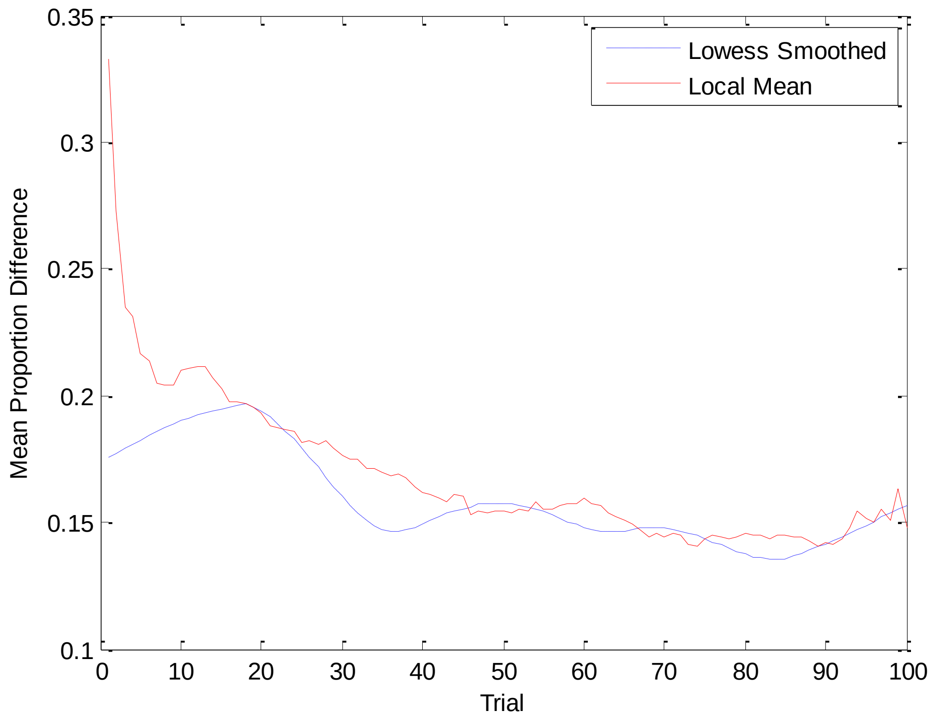

Figure 2.8 Mean size of the difference between estimates to the same items when seen in the first and second half of the experiment.

The critical items placed at the end of each half were then analysed: t-tests revealed no significant differences between the first and second estimates for any of the 10 items. Thus by the second half of the experiment there is no significant improvement or learning. So at least in this task, participants learning and adjustment to the decision environment reached (or is close to) an asymptote by 100 trials.

2.2.3.2. Rank Order Effects and Range Frequency Theory

RFT predicts that there will be a significant effect of the positively skewed distribution and that value estimates should be better predicted by an item's rank order within the stimuli set than by its target value. Mean estimates of value were first plotted against stimuli target values (Figure 2.9). This reveals what appears to be a good fit. However, one easily identified pattern is that lower value stimuli tend to be slightly (but reliably) over-valued while high value stimuli are under-valued. To test whether this is a deviation which can be explained by rank order, mean value estimates were calculated across participants for each item and plotted against their rank according to target values (Figure 2.10). This appears to be a worse fit. A correlation analysis confirms this as target values reveal a better fit to individual estimates ($r = 0.847$) than rank order($r = 0.835$). A Fishers z-test reveals that this difference is not significant ($z = 1.0$, $p>0.05$) but this is potentially due to the inherently high correlation between value and rank order ($r = 0.966$). The result also cannot be argued as a simple lack of power as it is actually in the opposite direction to that predicted by RFT. The theory would predict an equal spreading of estimates across the range of values. The pattern observed here is in fact in the direction of regression to the mean.

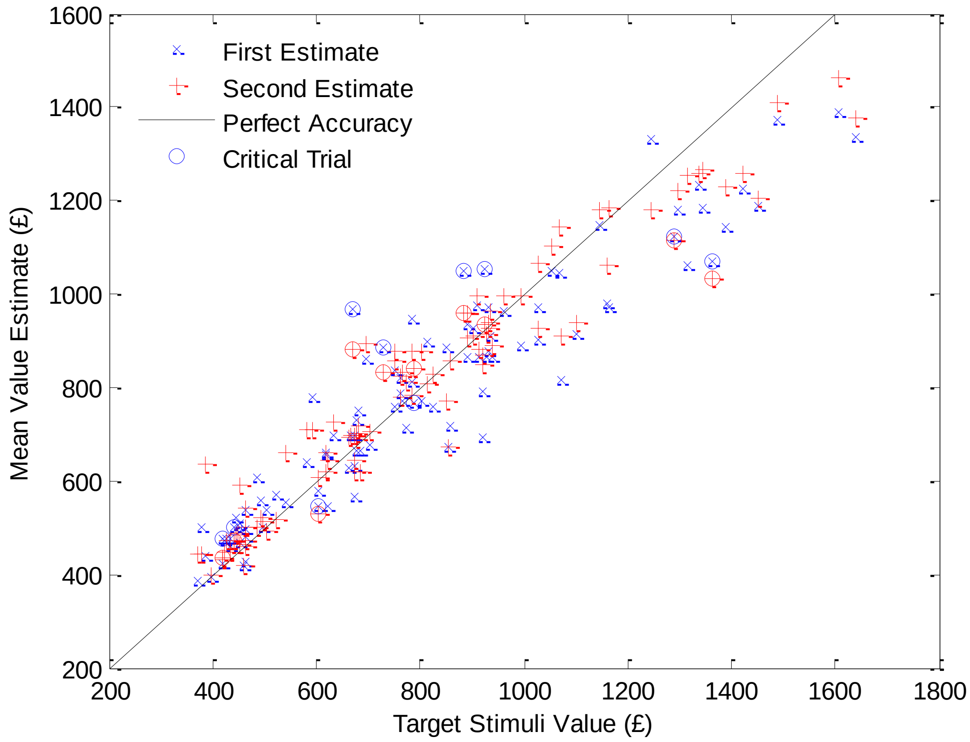

Figure 2.9 Relationship between mean estimates and target value

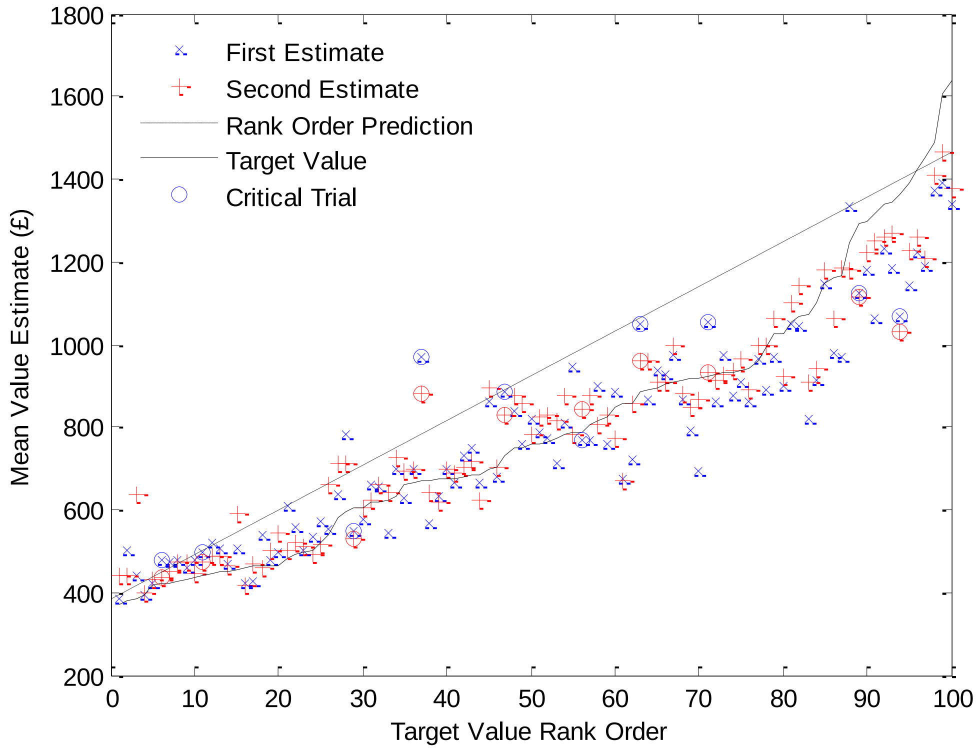

Figure 2.10 Relationship between mean estimates and rank order

2.2.3.3. Decision by Sampling

As DbS posits that decisions are based upon repeated comparisons at the attribute level, calculating its predictions must be done by finding each item's sum of favourable comparisons with other items and this must be done separately for each attribute. In order to calculate DbS scores, each attribute for each item was compared with every other item previously seen in the experiment. To control for the number of comparisons this score was then divided by the number of preceding items. The correlation with individual value estimates (r = 0.779, Figure 2.11) was weaker than the correlation between estimates and target values. A Fishers Z reveals the difference is significant (Z = 4.96, p<0.001), suggesting that DbS is a worse predictor than the true values participants are estimating.

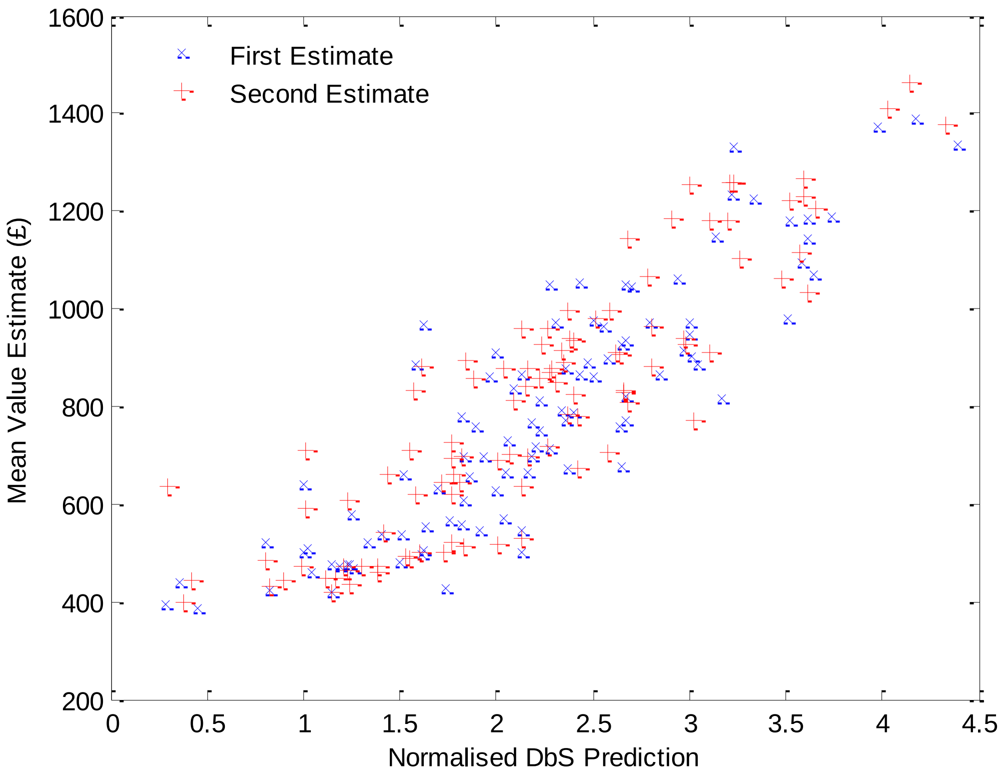

Figure 2.11 Predictions of DbS as it stands plotted against participants' mean value estimates

2.2.3.4. Decision by Sampling and Memory

Some aspects of DbS would seem to be implausible. The assumption that all previous memories, or at least a very large number of them, are sampled every time a new item is seen runs counter to evidence from both psychology and neuroscience. Is it the case that by imposing restraints upon the model to make it more plausible, it can also become more accurate? The first step is to control the number of previous items used in the comparison, i.e. limiting the sample size. To achieve this a cognitive model was created which simulated DbS but with constrained sample sizes. The mechanism of sampling was entirely random, with any previous attribute of any previous item having equal likelihood of being sampled and compared to the current one. The only constraint was that there was no re-sampling, meaning that for cases where the required sample size was larger than the number of previous attributes, all previous attributes were used. This model was used to create DbS scores for each item for sample sizes between 1 and 1000. For each sample size 1000 iterations were performed and each of the predicted values from these iterations was correlated with individuals' estimates. Then r-values were averaged across iterations to give a reliable estimate of the model's predictive accuracy.

If DbS is fundamentally correct in all but sample size then there would be a peak in model fit at a reasonably small sample size before a decline as the sample size becomes implausibly large. However, as seen in Figure 2.12 there is no such peak in fit when model predictions are correlated with participants' estimates. The fit is very poor for smaller samples and increases with an asymptotic shape, showing the strongest fit with an implausibly large 1000 item sample. This sample size makes the model equivalent to unmodified DbS and as such no improvement can be found upon the original model. The same asymptotic pattern is also found in DbS's correlation with target value, and this correlation is actually stronger. Thus it seems that the model's predictive ability increases as more items are added only because this results in a stronger correlation with target value, not with participants' estimates.

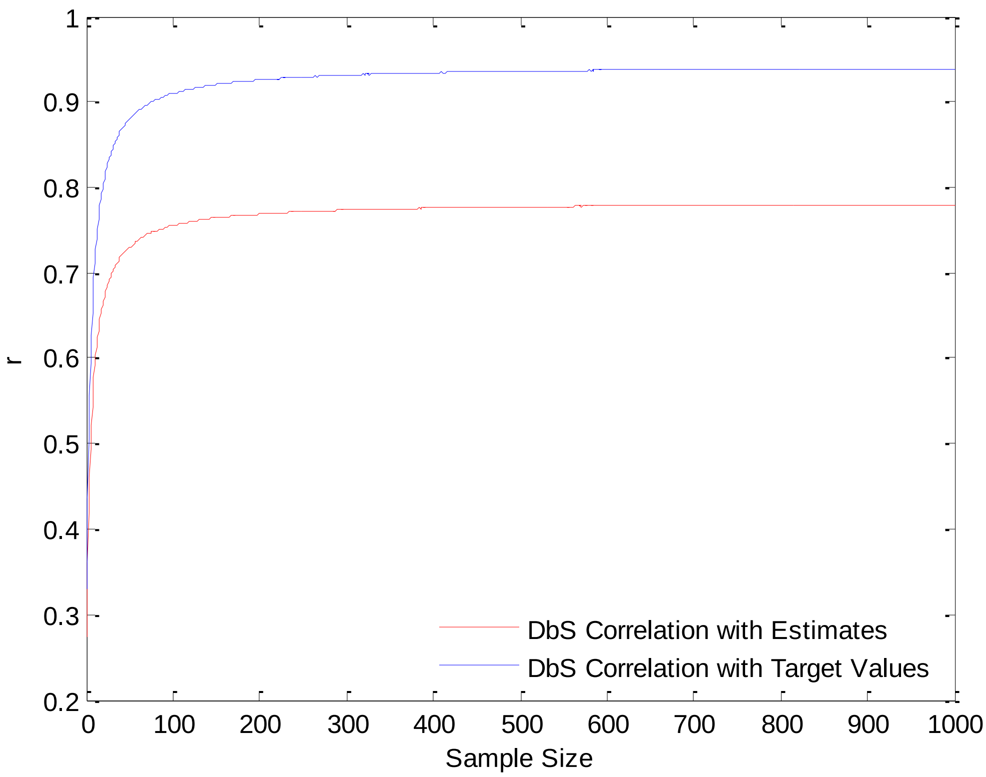

Figure 2.12 The strength of correlation between DbS predictions and either target values or individual estimates when DbS is modeled using varying numbers of comparisons with previous attributes

Although constraining the sample size makes the model more psychologically plausible, this is only one questionable assumption. The prediction that a sample is draw entirely at random from all previous experiences still ignores a vast wealth of findings from memory research. It is well established that during recall, there is a significant advantage for information seen most recently (Ebbinghaus, 1913) and information seen earliest in the task or block (Glenberg et al., 1980). The former is the recency effect and can be represented using the Ebbinghaus forgetting curve. This curve represents a rapid decay function whereby more recent experiences are most likely to be recalled and then decay quickly. In addition to the recency effect is the advantage for early experiences, i.e. a primacy effect. These two phenomena combine to form a U shaped serial order position curve. In this case, the earliest and most recent experiences are most likely to be recalled, with those in between being the least likely. There is evidence that primacy only occurs when participants know there will be a subsequent memory test and therefore make a conscious effort to memorize items (Marshall & Werder, 1972). Therefore, the effect of recency and primacy were modelled separately. The Ebbinghaus forgetting curve can be represented by a simple exponential decay function (Equation 6), with only memory strength left as a free parameter. Primacy can be modelled using the same equation, the anchor point is simply moved from the most recent item, to the first one encountered. The serial order position curve is the summed effect of the two exponential functions, one anchored at the first trial and the other anchored at the most recent.

Equation 6 The Ebbinghaus forgetting curve. R is retention, or in the case of DbS the probability of being sampled, t is the time since the item was experienced and s is the memory strength.

$$R = e^{-\frac{t}{s}}$$

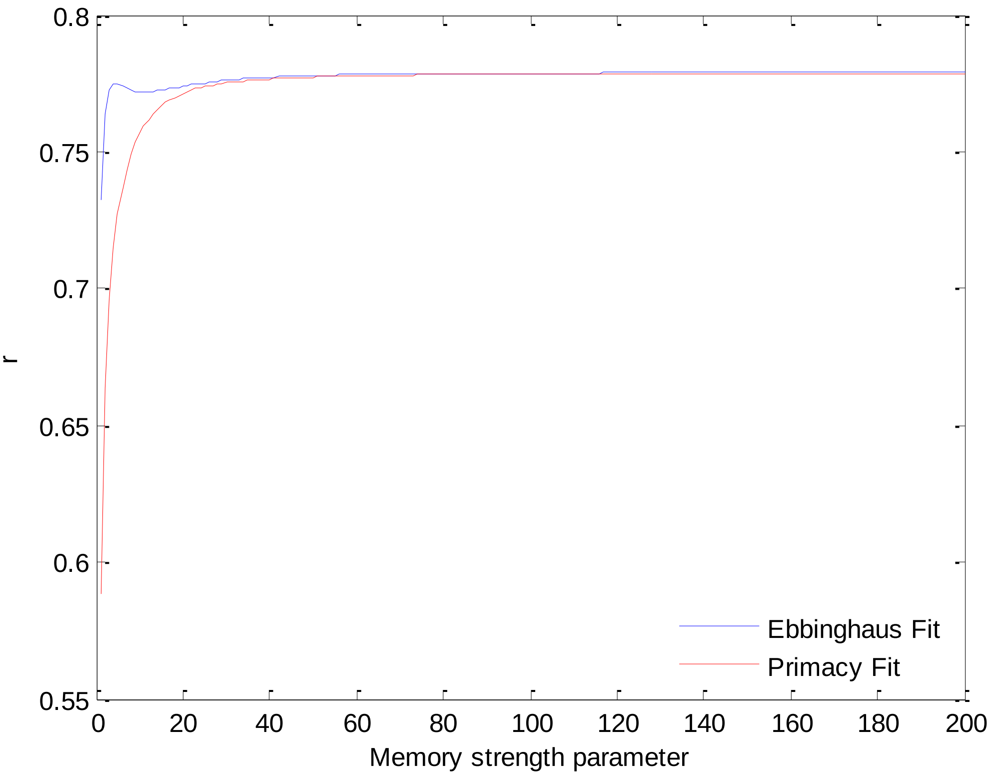

Figure 2.13 The correlation between value estimates and modified versions of DbS. Models weighted for recency and primacy were computed for memory strength parameters between 1 and 200

DbS was modelled using Ebbinghaus curve weighting functions for recency and primacy. If the current item compared favourably on a dimension then rather than incrementing a counter by one, it was incremented by the sampled item's associated weighting function value. Two separate model versions were computed: one for recency and one for primacy. For each of these, DbS was modelled with memory strength parameters between the values of 1 and 200. At lower values the weighting function is very sharply curved and the impact of an item decreases very quickly the further it is from the anchor point. At higher values, the curve becomes far less extreme and by 200 is close to linear.

As can be seen in Figure 2.13 the primacy weighting function has an asymptotic shape with fit approaching a maximum of $r = 0.78$ as the weighting curve becomes more linear (this is the same fit to two decimal places as is found when using no weighting function). However, the fit for the Ebbinghaus curve is very different at low values. It shows a local maximum at $s = 4$ before declining and then becoming very similar to the asymptotic shape found for primacy, as the increasing linearity of the functions inherently makes their predictions more similar. This is initially encouraging, but examination shows that it is only a local maximum ($r = 0.7749$) and still lower than that found using a weighting function of 200 ($r = 0.7789$).

When recency and primacy are applied at the same time a serial order position curve is created. The best fitting parameters were found using maximum likelihood estimation combined with annealing to avoid problems of local maxima. The results are very similar to that found for the separate weighting curves, but with the early maximum for recency now providing the best fit: recency = 3.02, primacy = 172.0, $r = 0.78$. Despite this, if we plot the fit against recency and primacy then the performance of the model is actually incredibly similar for the majority of the tested parameters (Figure 2.14). Therefore it seems very unlikely that the addition of this weighting curve will ever be able to solve the inherent problems of the model.

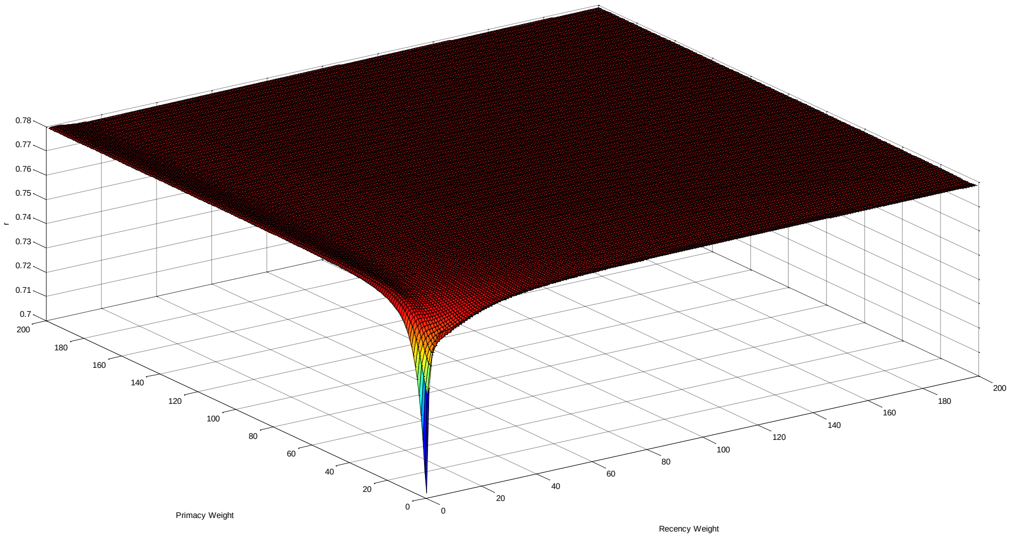

Figure 2.14 Correlation between value estimates and DbS predictions when weighted using a serial order position curve. The strength of correlation is shown at each value between 1 and 200 for memory strength on both primacy and recency

2.2.3.5. Decision by Sampling and Attribute Weighting

As well as assuming that the relative recency of experiences has no impact upon the likelihood of an experience being sampled, DbS also assumes that all attributes and different types of information are equally important. However this doesn't seem plausible; when considering apartments for example, it is highly likely that individuals will place more emphasis and importance upon attributes such as the number of bedrooms than they will upon whether they like the colour of the walls. By performing a multiple linear regression using the separate DbS scores of each individual attribute, the relative impact of each attribute can vary freely within the model. Furthermore a separate regression model was computer for each participant to allow for individual differences in weightings. DbS scores for each attribute are then multiplied by their calculated weight for each participant allowing a cluster corrected correlation to be performed to calculate the r-value.

Allowing attribute weightings to vary freely results in DbS showing a significantly better fit to the model than simply regressing the target values ($r = 0.88$, $z = 3.47$, $p<0.001$). However, the model now has a large number of free parameters. There are a number of statistical tools which could be used to control for the effect of these additional parameters but the design of the task lends itself to a mode parsimonious and valid approach. The same analysis was performed but using stimuli values instead of DbS scores. This controls for the number of free parameters and also provides a direct test to the Weighted ADDitive model which is regularly used as a baseline measure in multi-attribute tasks as it is analogous to multiple linear regression (Dieckmann, Dippold, & Dietrich, 2009). This reveals that the predictive accuracy of the two models are not significantly different ($r = 0.89$, $z = 0.37$, $p>0.05$).

2.2.4. Discussion

The results show no support for either RFT or DbS. Mean estimates correlated more strongly with target value than with the rank order predicted by RFT. Furthermore, DbS was significantly worse than a simple baseline in predicting participants' value estimates. This was true even when several modifications were made to the model to incorporate patterns of memory recall. The results for all but one version showed a trend towards worse predictions than baseline. The only version of the model which showed an improvement over simply correlating estimates with target values had a far greater number of free parameters. This model was no better than a baseline measure which was matched for degrees of freedom: simply regressing stimuli values.

The results shown in Figure 2.9 do suggest a systematic deviation from target values, with higher values being underestimated and lower values being overestimated. However, these deviations cannot be explained by RFT or DbS. Simple regression to the mean seems a more parsimonious hypothesis. It seems sensible that when unconfident, participants anchored their estimates closer to the mean of the distribution. This would also account for the reduction of the effect in the second half of the experiment, when participants were better calibrated.

A potential criticism of the experiment is high correlation between rank order and target values and also between DbS scores and target values. The distribution of values was created to match the positive skew found in many real world environments, but it does not allow for easy separation or orthogonalisation of DbS and target values. It could be that when the distribution of values is particularly novel, and significantly different to target values, these models perform relatively better. However, the finding that RFT and DbS performed worse does suggest that their poor performance in this experiment was not a ceiling effect.

2.3. Experiment Two

2.3.1. Introduction

As DbS and RFT are inherently reliant upon the distribution of values an individual has experienced, their ability to predict value estimates should remain stable as the distribution that the individual encounters changes, whereas the performance of the baseline of target values should decrease. Therefore Experiment Two uses stimuli drawn from a non-normal distribution. If value estimates are predicated upon the distribution and rank order of previously experienced values then the models' performance should remain relatively high whilst the correlation with target values declines significantly. This design directly tests a prediction drawn from RFT and increases the difference between predictions from RFT and target value compared to Experiment 1. Therefore, the experiment should provide a more powerful test. In addition, the 10 critical items were again presented at the end of each 100 trials. Because they remained the same on each presentation, their value did not change, but the difference in the distribution of preceding values meant several had a different rank. Thus RFT predicts that these items will elicit significantly different valuations on each presentation.

2.3.2. Method

2.3.2.1. Participants

Thirty-two members of the University of Nottingham community were recruited and were paid £5 inconvenience allowance. Their average age was 19.7 (S.D = 2.6). The sample size is larger than in Experiment one so that there is sufficient power for between subject analyses on order effects.

2.3.2.2. Stimuli

The same apartment stimuli were used as in Experiment One; with the same equations and weighting functions. However, instead of repeating the same items twice in each half of the experiment, the 90 original stimuli were manipulated to create a non-normal distribution. In order to create a portion of the distribution which was significantly under-represented, the top 40% of stimuli were taken and their values increased as a proportion of the distance from the most valuable item. This essentially results in the top 40% of the distribution being “squashed” up into the top 20%. The next lowest 5% of stimuli were then increased in a similar way so that that they occupied the under-represented portion of the distribution. This results in a broadly bi-modal distribution (Figure 2.15), which has the same rank order as the original distribution.

The stimuli values were changed whilst maintaining the relative influence of individual attributes between distributions. This was achieved by taking the outcome of the weighting functions of all continuous attributes for each stimulus and multiplying this by the third root of the ratio between original and modified rental vales. The resulting values were then entered into the reverse of equations 3,4 and 5 in order to calculate new attribute values on these scales.

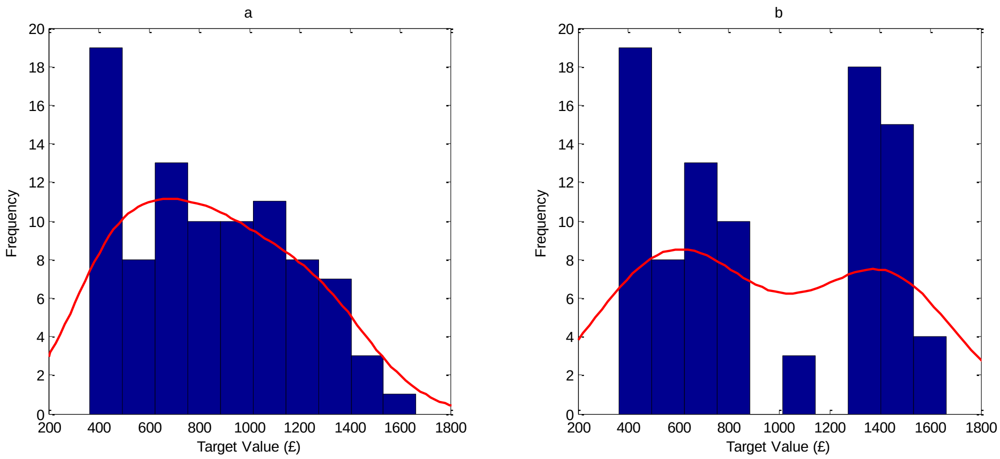

Figure 2.15 a. The original distribution continuous and with a slight positive skew. b. The modified distribution, non-normal and bi-modal.

2.3.2.3. Procedure

The same procedure was used as in Experiment One. The only difference was in the stimuli values presented to the participants. Half of the participants saw the stimuli drawn from the original distribution in the first and those drawn from the modified distribution in the second half. The rest of the participants saw the same stimuli, but with the block order reversed.

2.3.3. Results

2.3.3.1. Learning and Accuracy Rates

To confirm that participants were still able to learn from the feedback and calibrate to the stimuli values despite the altered distribution, error percentages were calculated in the same manner as in Experiment 1 and plotted by stimuli order. The data was separated by condition and Figure 2.16 shows that participants learn the stimuli values quickly regardless of whether they first saw the original or modified distributions.

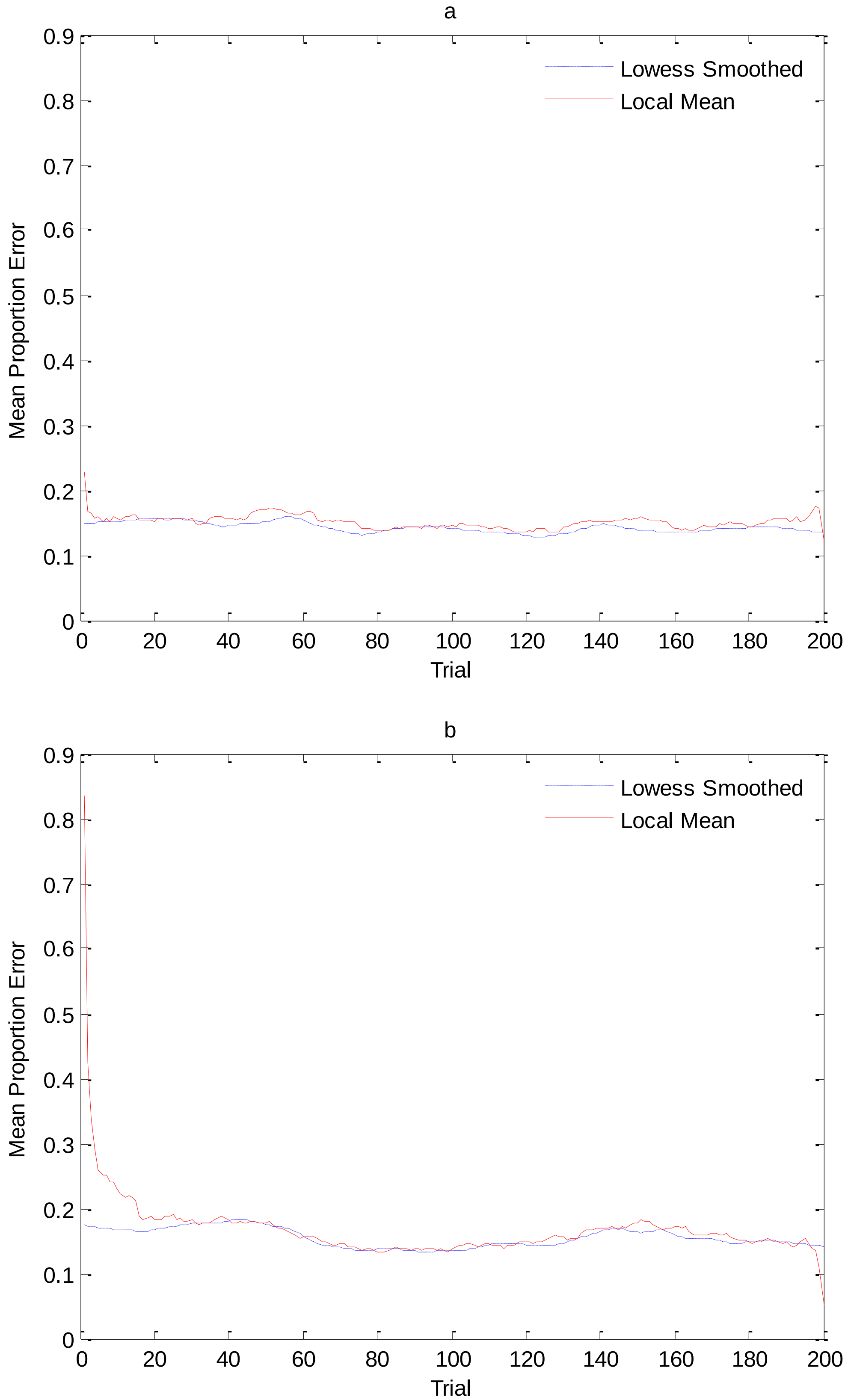

Figure 2.16 The percentage error rate over the course of the experiment for a. participants who saw the original distribution first and b. participants who saw the bi-modal distribution first

2.3.3.2. Rank Order Effects and Range Frequency Theory

A 2(condition)X2(distribution) mixed model ANOVA was performed on each of the critical values placed at the end of each block. The stimuli and their values are identical at each presentation but for those with a value above £886 the change in distribution means their rank order is significantly different. Therefore RFT predicts a significant main effect of preceding distribution upon those with higher values but not for those with lower. Four of the ten critical stimuli do show a main effect. Three of these have higher target values and therefore have a significantly changed rank (£911, $p = 0.022$; £1291, $p = 0.045$; £1374, $p = 0.016$), but one is a lower value item where no difference is predicted (£413, $p = 0.036$). One must be careful of interpreting these results though, as none of these effects are large enough to survive correction for multiple comparisons. However what is particularly interesting is that all of these differences are again in the opposite direction to that predicted by RFT. Critical values have a lower rank within the modified distribution than the original but average value estimates are in fact higher. Curiously, this serves to keep the items' rank by value estimate stable between conditions. This is despite the target value being different.

The overall correlation between target values and value estimates in the second half of the experiment was then calculated. This was $r = 0.79$ for those who saw the modified then original distribution and $r = 0.87$ for those who saw original then modified (Figure 2.17). Both showed the same strength correlation with rank order ($r = 0.79$ and $r = 0.87$ respectively). This suggests that despite the unusual distribution, rank order and target values are equally good predictors of value estimates.

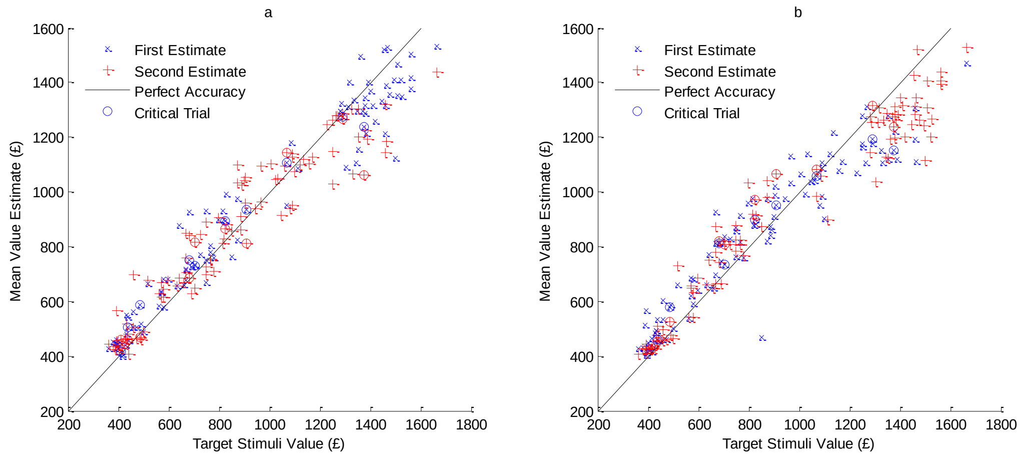

Figure 2.17 Relationship between target values and mean value estimates for a. participants who saw the original distribution first and b. participants who saw the bi-modal distribution first

2.3.3.3. Decision by Sampling and Memory

When DbS is explicitly modelled it again shows very slightly worse performance than simply correlating target values (modified then original, r = 0.75; original then modified, r = 0.85). There also appears to be surprisingly little difference between the DbS scores of items in the two distributions (Figure 2.18). This is largely because the model uses comparisons at the level of the attributes in order to make its valuations, meaning the rank order effect of the overall value is diluted. In addition, the model as it stands gives equal weight to all previous experiences, including those in the previous half of the experiment where the other distribution of values was experienced. Therefore one would expect a more significant improvement of the DbS model once recency is added as a parameter.

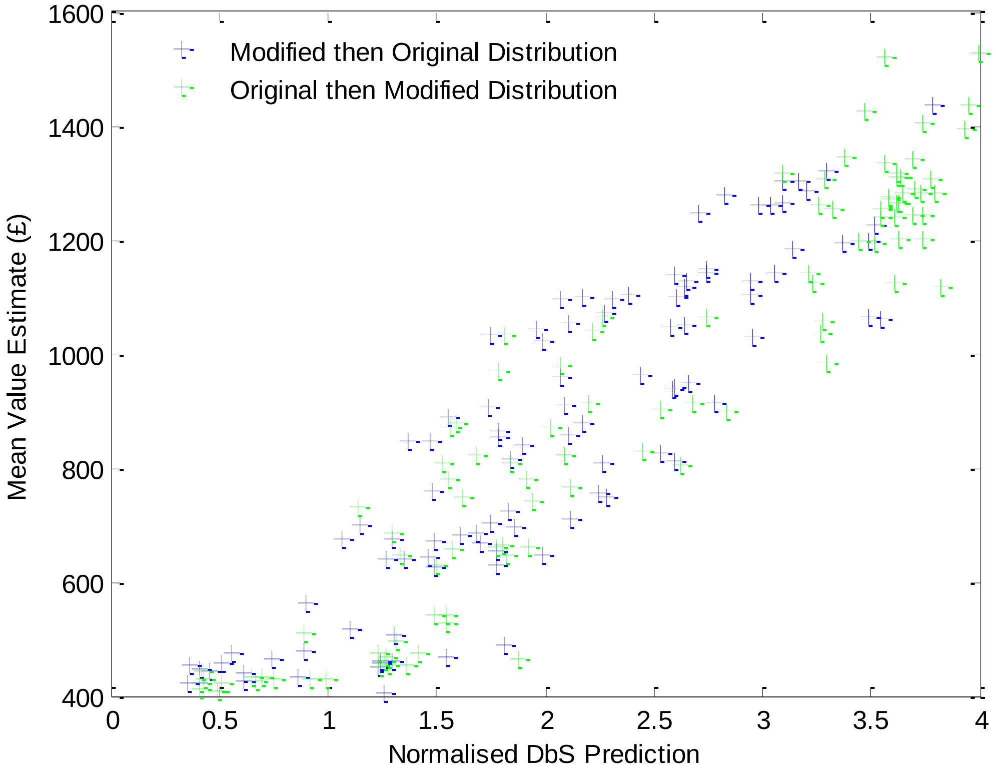

Figure 2.18 Relationship between DbS predictions and average value estimates in the second half of the experiment only

When DbS was modelled with recency and primacy parameters, no improvement was found for either (Figure 2.19). All r-values were the same as the un-modified model to 2 decimal places. For the original then modified condition the best Ebbinghaus fit was found at s = 119 (r = 0.85), and for primacy the best fit was found at s = 200 (r = 0.85). For the modified then original condition the best Ebbinghaus fit was found at s = 53 (r = 0.75) and for primacy the best was found at s = 200 (r = 0.75). Recency and primacy were then combined to form a serial order position curve. Maximum likelihood modelling then identified the best fitting parameters. For original then modified, primacy s = 0.0072 and recency s = 119.34 produced the best fit. For the modified then original condition the best fit was found at primacy s = 1.678 * 10^14 and recency s = 99.7. The latter in particular seems unrealistic and when the fit of the model at different parameter values is plotted the same pattern is found as in Experiment One (Figure 2.20). The model performs poorly when both parameters are small, but then shows a rapid increase before performing incredibly similarly for all other values tested.

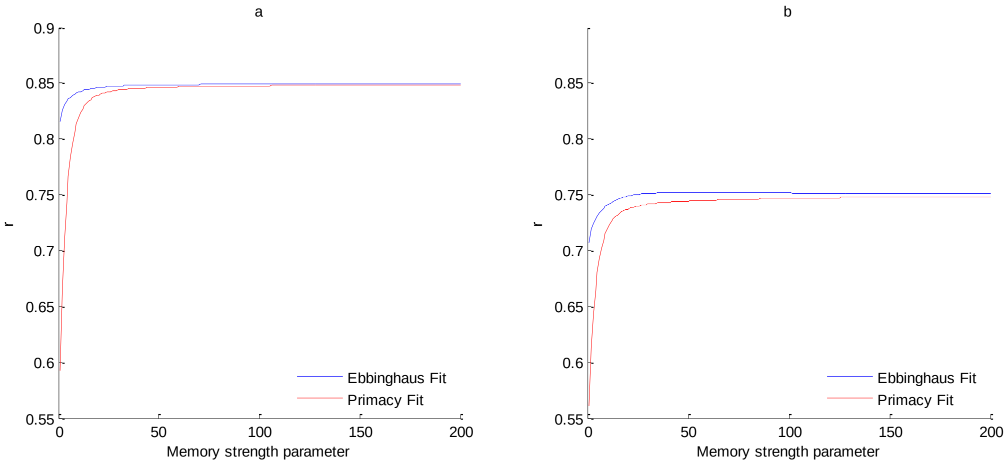

Figure 2.19 Correlation between value estimates and modified DbS models using memory strength parameter values between 1 and 200

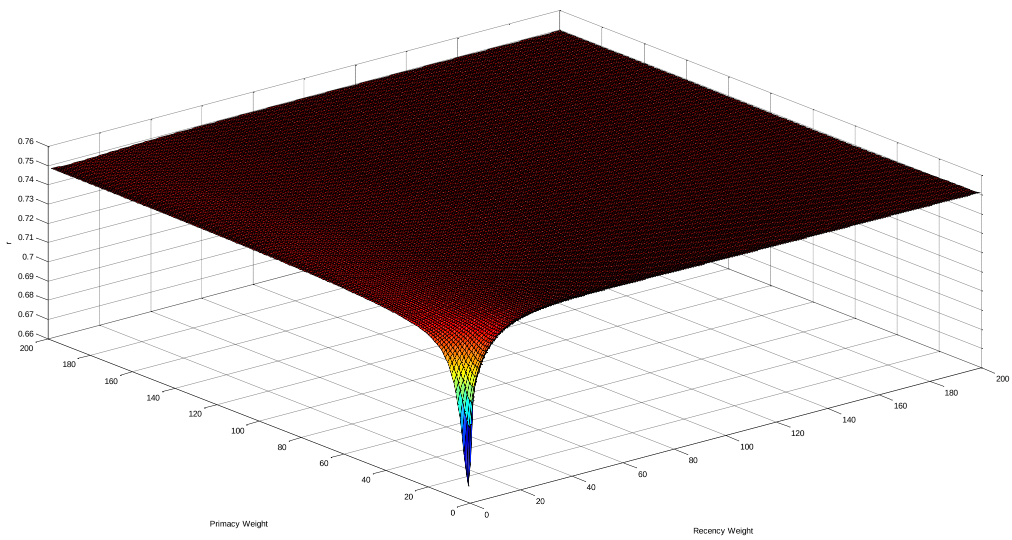

Figure 2.20 Modified Distribution First

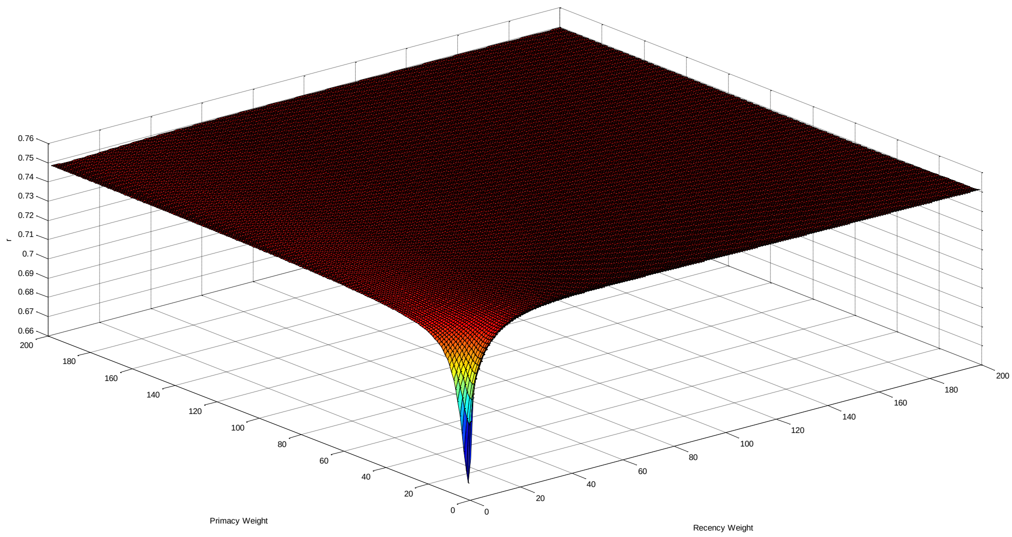

Figure 2.21 Original Distribution First

2.3.3.4. Decision by Sampling and Attribute Weighting

The same series of regression equations were used as in experiment 1 in order to estimate participants' attribute weightings and then applying these to DbS scores in order to compute predicted valuations. This again showed better performance than simply regressing true values, for both modified then original ($r = 0.90$, $z = 4.92$, $p<0.001$) and original then modified ($r = 85$, $z = 4.98$, $p<0.001$). However, when compared to the same model using stimuli values there was no difference in accuracy for modified then original ($r = 0.85$, $z = 0.06$, $p>0.05$) or original then modified ($r = 0.91$, $z = 0.81$, $p>0.05$).

2.3.4. Discussion

This experiment modified the distribution of experienced values from the first to the second half of the task. Because both RFT and DbS are predicated upon the distribution of experiences and comparison with recent items, their predictions change when the distribution does. Therefore the models should perform better than the baseline measure of simply correlating estimates with target values. However, when the distribution changes, the performance of the DbS model declines at the same rate as that of the baseline measure.

The main finding from these two experiments is that the performance of RFT and DbS rests upon their high correlation with target values. When the correlation between target values and participants estimates becomes better or worse, there is a concomitant change in the performance of both models. Therefore the simplest conclusion is that these comparative models do n ot explain deviations from perfect accuracy. However, it is suprising that target values and rank order have virtually identical predictive accuracy in each presentation order. This is despite the two making significantly different predictions. Thus, there remains the possibility of an interaction between the two but an inability to rule out the possibility should not be mistaken for evidence in favour of it.

One interesting finding is that, although not surviving correction for multiple comparisons, a number of the critical items suggest a difference in value estimates depending upon the preceding values. What is particularly interesting is that these are in the opposite direction to that predicted by RFT. Rather than a higher rank order by target value resulting in a higher valuation, average estimates become smaller. This higher value estimate in the modified distribution results in the items' rank *by value estimates* staying relatively similar. Estimates rise relative to those in the original distribution along with that of other surrounding stimuli whose target value is also higher. This could be the result of a simple regression to the distribution mean, or using the mean of recently viewed values as an anchor. Hence when, as in the modified distribution, there are more values clustered around the higher end of the range the mean is higher and so the value estimates become higher. Alternatively, the increased number of items with higher values could simply be extending the perceived range of stimuli values. If this is assumed then the results would fit better with RFT.

2.4. Chapter Discussion

This chapter details two experiments which assess the ability of RFT and DbS to predict responses on a multi-attribute valuation task. Cognitive modeling techniques were also employed to assess potential modifications to the model inspired by findings from the memory research literature. In summary, the results do not support either RFT or DBS. The models closer to the original models, with fewer free parameters, were consistently outperformed by a simple baseline measure. The only models which were not significantly worse than baseline were versions of DbS with a large number of free parameters. However, the performance of these models was not significantly different to that of a WADD model when matched for free parameters.

The modifications to DbS which were tested were an attempt to reconcile the model with findings from memory research whilst also making the model more neurologically and psychologically plausible. The most basic modification was to restrict the number of previous experiences sampled when calculating value. However, rather than reveal an improvement at a plausible sample size, the performance of the model improved asymptotically as the sample size increased. The same relationship was found between model predictions and target values. Thus as the model used a larger proportion of items to calculate values, the resulting DbS scores became more similar to the target values. This was a pattern found throughout all analyses and it seems the model becomes better at predicting estimates only because it gets better at predicting the target values. If the model were predicting behaviour rather than the original stimuli values then its performance would not have such a reliable correlation with target values and would surpass the predictive abilities of those target values.

Perhaps the most consistent finding within memory research is that more recent experiences are more likely to be recalled. Therefore an Ebbinghaus forgetting curve was added to the model as a weighting function and a large range of curvature parameters tested. Despite the large number of values tested, the modification failed to improve performance. The data suggested a local peak in performance when using a weighting function with a steeper curve in Experiment One, but this still performed poorer than an unmodified version of DbS and no such peak was found in Experiment Two. Apart from this minor deviation, the shape of the fit was broadly asymptotic; with the model performing better as the weighting function became more linear and was therefore using more information.

In addition to recent items being more frequently recalled, many experiments also find an advantage for the first few items experienced. Applying a primacy weighting function revealed an asymptotic relationship between linearity of the function and the performance of the model, similar to that found when examining sample sizes. This is likely because the task was not explicitly described as a memory test and therefore the lack of rehearsal meant there was no primacy effect (Marshall & Werder, 1972). By combining both recency and primacy into a single weighting function a serial order position curve was created, but even with these additional free parameters DbS did not achieve the same predictive accuracy as a simple correlation between estimates and target values. Furthermore, plotting the performance of the model against parameter values reveals that performance is very poor when both parameters have small values (so extreme curvature), but quickly asymptotes and has a very flat fit for all other values modelled. Substantial changes in parameter values result in little or no change in performance. This is a good indication that such a version of the model is inefficient or inaccurate, with very high variance in parameter estimates (Busemeyer & Diederich, 2010).

As stated in the introduction, DbS is a model of decision making rather than valuation. One must therefore assume a mechanism for transforming a score from an internal psychological scale into a financial value. However, such a transformation cannot be driving the poor fit of the model here. If this were the case then value estimates would still have to increase monotonically with DbS scores. This is also true for a range frequency model. The shape of the relationship would not necessarily have to be linear but any transformation complex enough to violate monotonicity would necessitate an entirely different model. Plotting predictions against value estimates revealed no such pattern.

In Experiment two critical trials with values in the modified portion of the distribution received significantly different valuations depending upon the distribution of preceding values. When the distribution had been modified so that higher values were over-represented and mid-range values under-represented, value estimates for the higher value control items were higher. One potential explanation is that participants are simply using a method of anchor and adjustment (Tversky & Kahneman, 1974). When control items are viewed after the original distribution then the average of recently viewed items is lower than it is if the modified distribution immediately precedes them. Therefore anchoring to a higher or lower average, results in the higher and lower estimates observed here.

Another possible explanation is adjustment to the range of the values experienced. In the original distribution high values are under-represented, and as a result participants could perceive the range of values to be lower than in the modified distribution. Participants could potentially be adjusting to the perceived increase in range by increasing their estimates (although it should be noted that the actual range of estimations varies very little). Furthermore, many stimuli receive value estimates in the mid-portion of the range despite being significantly under-represented in the true stimuli values. This suggests an adjustment according to the frequency of estimates at different value ranges. These two phenomena are predicted by RFT (Parducci, 1965). The increase in perceived range results in the estimated value of critical items rising, whilst the uneven frequency distribution of experienced values results in a significant number of estimations within the under-represented portion of the range.

In both experiments, participants performance was remarkably high and had very high correlation with target values. Therefore it seems participants are able to incorporate feedback and develop weighting functions when there are more than two cues (Todd & Hammond, 1965). There are several likely reasons why participants in these studies were able to whilst those in others were not. Firstly, participants begin the experiment already aware of the valence for each attribute and had reasonable expectations regarding each one's relative impact based on real-world experience. This is analogous to having task information at the beginning of the task, which is known to dramatically improve performance (Balzer et al., 1989). In addition, the relationship between cues and outcomes was entirely deterministic. The majority of studies where individuals have not shown learning have been probabilistic and contained random noise (Harvey & Fischer, 2005). Some argue that individuals are not able to properly integrate probabilistic outcomes and cannot separate the random noise from the signal as would be required to abstract weighting functions (Brehmer, 1980). Our results support these previous findings: in an entirely deterministic task where participants do not have to separate signal from random noise, they perform with high accuracy.

In chapter three the paradigm is modified to examine the effects of individual attributes and their distributions. The experiments directly address which has the greater effect upon value estimations: the true mathematical weighting function linking attributes and rental value, or the distribution of experienced attribute values. This detailed examination of attribute use also allows a comparison between heuristic accounts of valuation and mathematically compensatory accounts such as WADD.

3. Chapter 3

3.1. Chapter Introduction

In this chapter two experiments are reported which test the predictions of DbS by direct experimental manipulations. In the previous chapter the distribution of rental values was manipulated. This tested the predictive accuracy of RFT in multi-attribute tasks as the model makes predictions about how the value of an item is transferred from an internal psychological scale to an explicit financial valuation. However, DbS predicts that the same rank ordering strategies occur in the production of the internal psychological valuation, at the level of the attribute. Therefore the modified distribution of rental values resulted in very little change in the predictions of DbS. In this chapter, two experiments were conducted which again use multi-attribute valuation tasks. In each task two of the items' attributes use the same scale; however one has a modified distribution, whereas the other has a modified weighting function. Thus, by extracting participants' use of these attributes and their relative weighting, it is possible to directly test DbS.

Experiment 4 also addresses the issue of prior expectations. Experiments 1-3 have used apartments on the basis that the undergraduate population would have limited experience with specific examples. However, they will undoubtedly have pre-conceptions and some experience of values from their life up until that point. These more distant experiences were predicted to have minimal effects during the task based upon memory phenomena (Ebbinghaus, 1913) and previous results in choice and valuation tasks which show large effects of immediately preceding items (Beckstead, 2008; Stewart, 2009; Ungemach et al., 2011; Vlaev & Chater, 2007). But the results of Experiments 1&2 did not support this. Therefore it is possible that values experienced prior to the experiment still had a significant effect upon value estimates and this could be the reason the distribution of values did not modify valuations in the manner predicted by DbS and RFT. Experiment 4 addressed this by using exactly the same stimuli values, but with a very different cover story. By asking participants to estimate the values of mineral deposits based on levels of contamination etc.

it was possible to compare the use of information between situations with (some) existing knowledge and those with none.

The detailed examination of attribute use and weighting functions in this chapter also allows the results to speak to one of the most fundamental debates within JDM research: Whether individuals use mathematical functions and compensatory weighting of information, or rely upon simpler heuristics and rules (Gigerenzer & Selten, 2002). Although most research into heuristics has focussed upon choice between alternatives (Brandstätter et al., 2006; Bröder, 2002; Glöckner & Betsch, 2008; Johnson, Schulte-Mecklenbeck, & Willemsen, 2008; Newell & Shanks, 2003), there are proponents of heuristics in single item valuation (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 2000). Furthermore, the experiments elicit explicit estimates of subjective importance of different attributes and information. Thus, if the results support a Weighted ADDitive (WADD) account, it is possible to examine whether people's stated estimates of attribute importance are an accurate reflection of their actual use and weighting of the information.

3.2. Experiment Three: Distributions and Weighting Functions

3.2.1. Introduction

This Experiment used a very similar task to that of Experiments 1 &2 with participants estimating the rental value of a series of apartments. However, the stimuli were changed so that the predictions of DbS could be directly tested. DbS predicts that the value of each attribute for the current item is compared to previously experienced values on that particular attribute. Thus it is the distribution of values within an attribute which dictate the estimated value, not the distribution of the overall rental values. So if RFT predicts that rank ordering is used to transform values from an internal psychological scale to an external response scale, DbS predicts that the rank effects occur earlier, during the valuation on the psychological scale

The biggest change from the task in the previous chapter was that the number of bathrooms and land-lord rating were removed from the stimuli attributes. The bathrooms attribute was removed because it was not independent of the number of bedrooms. A property with 1 bedroom and 4 bathrooms is unrealistic, but constraining the number of bathrooms by the number of bedrooms means that there was significant correlation between the two attributes. The land-lord rating was removed because anecdotal reports from participants suggested it was the scale they found most difficult to interpret. It is also the only one which doesn't have a credible real world analogue. These two attributes were replaced by a crime risk rating. This was a score between 0 and 10 representing the risk of being the victim of crime in that postcode. By then constraining the maximum distance from town the values given to these two attributes could be interchangeable. It was then possible to manipulate the distribution of values for one attribute, whilst manipulating the mathematical effect of the other attribute upon rental value for the other.

The experiment also explored a more general issue in the JDM literature by attaining subjective estimates of attribute importance. Previous research has suggested that participants are generally quite poor at estimating their own strategies and there is little correlation between their subjective reports of information use and the information weighting statistically extracted from their behaviour (Cook & Stewart, 1975; Reilly & Doherty, 1989; Snook et al., 2011; Zhu & Anderson, 1991). However, these weightings are extracted by regressing stimuli values against valuations/choices. If individuals estimate values based upon rank order within attributes then it could be that the statistically abstracted weightings are simply calculated upon different information than that used by participants. This would explain why very different methods of extracting subjective estimates produce consistent measurements, but that these measured weights then do not correlate with those statistically extracted (Cook & Stewart, 1975). It is also a potential explanation for studies which find participants are far better at estimating a choice environment's weighting functions if they are those abstracted from another individual, rather than created by an abstract mathematical relationship (Reilly & Doherty, 1992).

A further advantage of measuring subjective attribute weightings is that it provides a weighting which is independent of the data. These weights can then be used in models of DbS where attributes are given different weightings. In Experiments 1&2 weighted models performed well, but weightings were estimated directly from participants responses, adding a large number of free parameters. By constraining the model using independently measured weights the model may potentially retain this high level of predictive accuracy without the increase in free parameters.

In testing the weighting functions used by participants there is an implicit assumption that they are using some form of mathematically compensatory system. However, these same analyses can provide evidence that a WADD model cannot properly describe the data (Busemeyer & Diederich, 2010). Thus, if participants make no significant use of any parameters or weight any single attribute highly enough that no combination of other attribute values can significantly alter the valuation, then these are strong indicators for a non-compensatory heuristic account (Brandstätter et al., 2006; Bröder, 2002; Gigerenzer & Selten, 2002; Gigerenzer & Todd, 2000; Glöckner & Betsch, 2008).

3.2.2. Methods

3.2.2.1. Participants

Thirty-two undergraduates at the University of Nottingham completed the experiment for course credit. Twenty were female and twelve male, with an average age of 20.4 (SD = 2.1).

3.2.2.2. Stimuli

The stimuli were 200 apartments. Four attributes were used in this experiment and their design differed from previous experiments in that they erred towards simplicity, sometimes at the expense of accurate representation of the real world. The variables used were number of bedrooms, size in square ft, distance from town and crime risk score. Number of bedrooms was a random integer between 1 and 4. Size was randomly selected between 750 and 2000ft using a boxcar function.

Distance and crime were the two variables of interest. Distance was given a linear distribution and cubic weighting; crime was given a cubic distribution and a linear weighting. Size was the only continuous variable which had both a linear distribution and a linear weighting upon rental value. Thus it acted as a control variable. The linear distribution of distance was created by drawing numbers from a boxcar function between 0 and 10, rounded to one decimal place. The cubic distribution was calculated by starting with a random number from a boxcar function between 0 and 10, rounded to one decimal place. The mean was then subtracted from each value creating an equal distribution around zero. This was then cubed before being linearly rescaled such that the smallest value was 0 and the largest 10.

Each variable's effect upon rental value was then calculated using the following equations

Equation 7

$$1 + \frac{(beds - 1)}{4} * 0.6$$

Equation 8

$$1 + \frac{(size - 750)}{1250} * 0.2$$

Equation 9

$$1 + \frac{(10 - crime)}{5} * 0.25$$

The effect of distance was calculated in the same manner as crime's distribution was, being centred around zero, cubed and then re-scaled to between 0 and 10 before being divided by its mean and multiplied by 0.25. Figure 3. shows the weighting function of each variable. The result is that both distance and crime have the same overall effect upon rental values, but the cubic and linear components are provided by either distribution or weight. Figure 3.2 shows the relative rank effects of the crime risk attribute. The cubic distribution results in DbS predictions of a steep increase in the centre of the scale with plateaus near the extremes of the range.

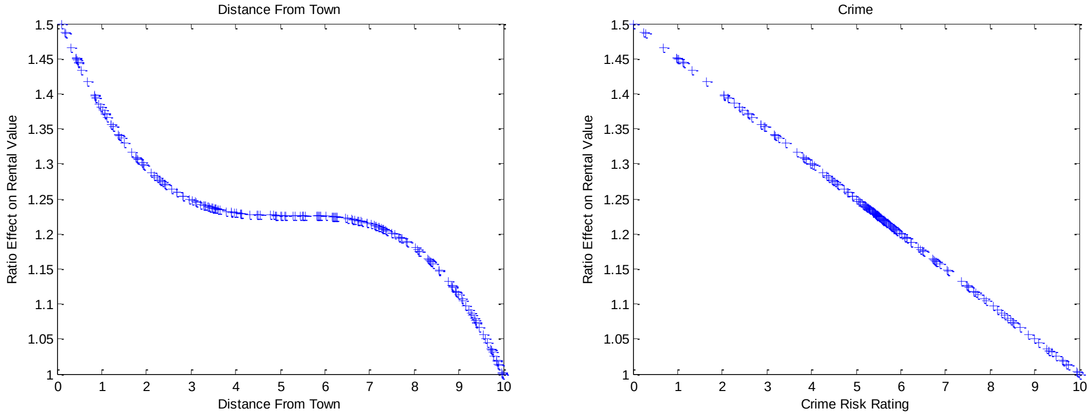

Figure 3.1 The effect of distance from town (left) and crime risk (right) upon rental value. Note also the different distributions, with crime being under represented at the extremes of the scale

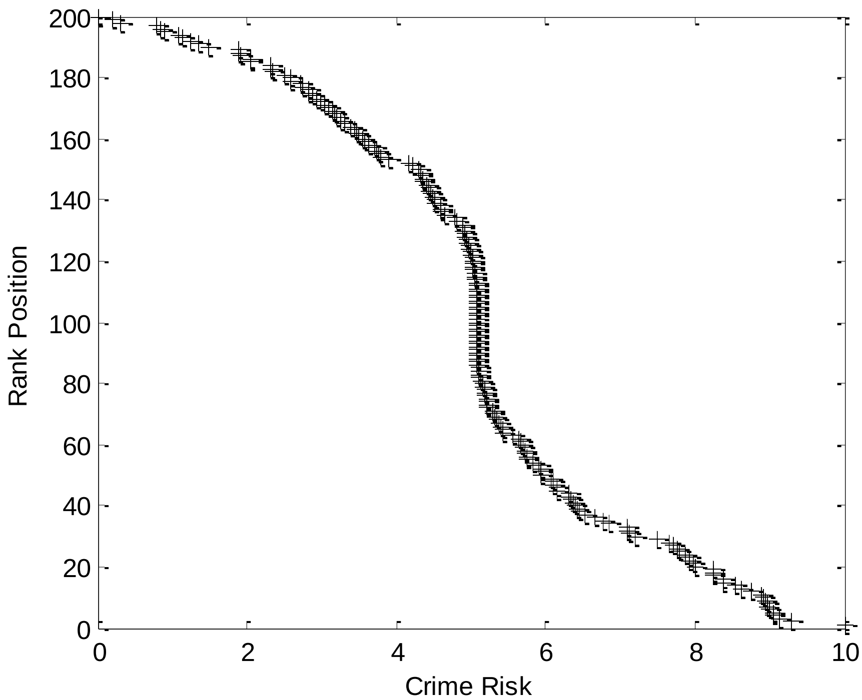

Figure 3.2 The relative rank effect of the crime attribute, i.e. the attribute's effect upon value estimates as predicted by DbS

3.2.2.3. Procedure

The procedure was similar to the previous experiments. Participants saw the details of an apartment and responded by typing their estimated rental value. They were then shown the true rental value for 2 seconds. There was then a 1 second interval of blank screen before the next item was presented. Participants were also asked to estimate how important each variable is when judging the value of an apartment. They responded using a slider on a scale of 1-100 labelled at either end as "not at all important" and "very important". Participants provided these ratings twice: once at the beginning and once at the end of the experiment so that it was possible to identify any effects of learning.

Participants were told that the task was part of an estate agent training programme. The crime variable was explained to participants as the crime risk in the property's post code, as calculated by the government website "police.uk", with a higher score indicating a higher likelihood of being the victim of crime.

3.2.3. Results

3.2.3.1. Accuracy and Learning Rates

To examine participants' accuracy and learning rates the average error rate was calculated for each item as a percentage of the item's true value. When these were plotted by trial, participants showed rapid learning and an asymptote at around trial 60 (Figure 3.3). Individual's responses from the second half of the experiment were then correlated with true rental values and show that participants are performing with high accuracy ($r = 0.708$). Despite the strong correlation, accuracy was significantly lower than in Experiment 2 ($z = 2.45$, $p<0.05$).

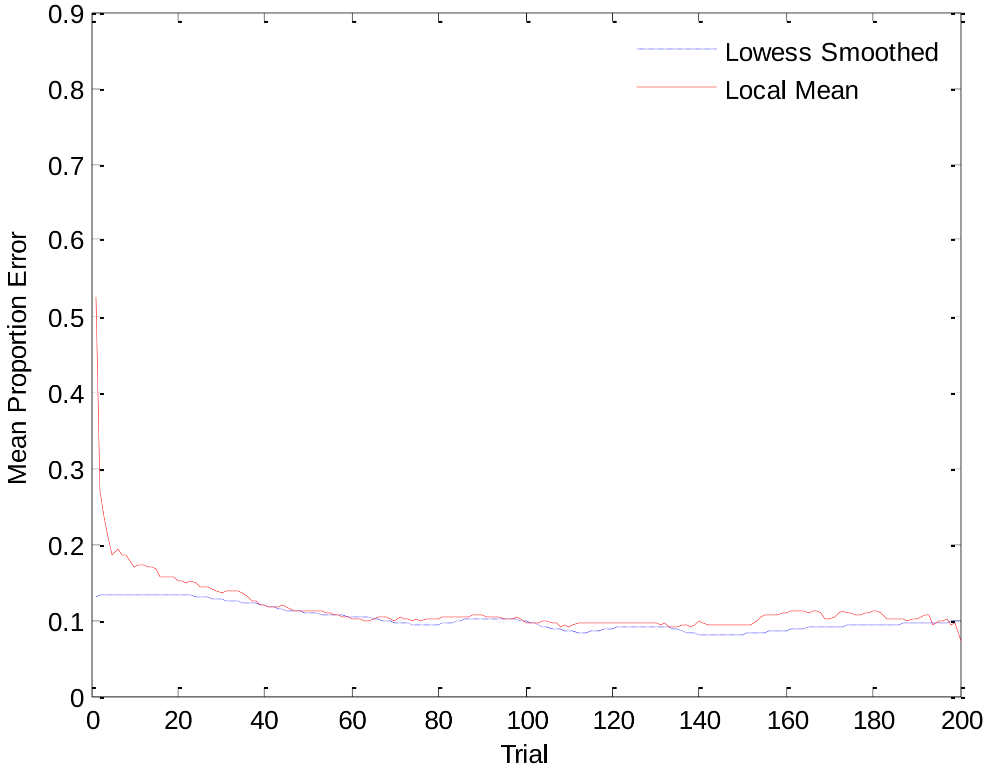

Figure 3.3 Average error rate during the experiment. Results are smoothed using a local mean and a lowess local regression.

3.2.3.2. Attribute Weighting Estimates

Participants provided estimates of each attribute's importance both at the beginning and end of the experiment. The aim was to investigate whether these estimates were an accurate representation of the true weighting functions that participants were trying to learn and whether they had insight into their own weightings. Standardised beta weights were first calculated by regressing the attribute values against the true rental values in order to serve as a measure of participants' accuracy with respect to the relationships they were trying to learn. Individuals' estimates were then used to extract participants' weighting of information when making estimates. A cluster corrected correlation then demonstrated that individuals were well calibrated when making their valuations and their use of information closely matched that of the mathematical relationship between stimuli and true values ($r = 0.83$, $p<0.0001$). There is a general pattern of participants underweighting all attributes, but the relative importance attributed to each attribute still demonstrates the significant correlation.

Participants' subjective estimates of attribute importance were analysed to examine their accuracy, starting with ratings provided before beginning the task. There was no significant correlation between participants estimates of attribute importance and the true weightings ($r = -0.17$, $p > 0.05$). There was a significant correlation between estimates and participants own revealed weightings but it was actually negative ($r = -0.26$, $p < 0.05$). This was primarily driven by participants underestimating their use of bedrooms and overestimating that of size (Table 3.1). In fact only 30.4% of participants correctly identified the attribute which they weighted most highly during their valuation estimates. Thus suggesting that participants are not using simple heuristics (Dawes, 1979) or prioritizing attributes in a hierarchical manner (Ayal & Hochman, 2009; Birnbaum & LaCroix, 2008; Brandstätter et al., 2006; Dhami, 2003).

A similar pattern is revealed by estimates provided at the end of the experiment. There is no significant correlation with true weightings ($r = -0.01$, $p > 0.05$). The relationship between estimates and subjective weightings was no longer significant ($r = -0.09$, $p > 0.05$), but the trend is still towards a negative correlation. The proportion of participants correctly identifying the attribute they weighted most highly fell to 21.7%, suggesting that participants' estimates of their own weightings did not significantly improve over time.

Table 3.1 The standardized betas when regressing stimuli values against true rental values, average estimates accross participants and the mean betas when regressions are performed separately for each participant. These are shown alongside the average subjective estimates of importance for each attribute.

	Truevals	Average separate regressions	Pre-task estimate	Post-task estimate
Intercept	660.2	650.4		
Beds	0.770	0.658	72.5	74.6
Size	0.312	0.174	82.1	71.6
Crime	-0.406	-0.143	77.0	90.2
Town	-0.412	-0.304	71.9	69.0

3.2.3.3. Decision by Sampling and Modelling

DbS scores were calculated by finding the number of previous stimuli which were worse than the current item on each attribute. This was then divided by the number of previous trials. Correlating these scores with individual value estimates for the second 100 items reveals that DbS is significantly worse at predicting valuations than simply correlating true rental values ($r = 0.603$, $z = -6.27$, $p < 0.05$).

A multiple linear regression was performed to allow the weights for each attribute to vary freely within the model. This revealed a fit of $r = 0.80$, which was the same as the performance when stimuli values were entered as predictors ($r = 0.80$). The experiment also allows examination of whether the weighted DbS model could retain its increased predictive accuracy when the values of the additional free parameters were independently measured and constrained. The DbS scores for each attribute were multiplied by participants' estimates of the attribute's importance. The resulting correlations were poor and slightly weaker than the original unweighted model of DbS. This was the case both for estimates provided before the task ($r = 0.55$) and after it ($r = 0.50$). This poor performance is unsurprising given participants' poor accuracy when providing these importance estimates (see above). As a further test of this assertion, the same model can be calculated using stimuli values rather than DbS scores. The importance weightings were divided by the standard deviation of their corresponding attribute scales before being multiplied by the attribute values and summed together for each item. This also revealed a very poor fit for weightings from before (r =

0.53) and after the task ($r = 0.45$). When the post-task weightings are used, DbS does significantly outperform the model using stimuli values ($z = 2.26$, $p<0.05$), suggesting that rank orders could be playing a larger role in participants subjective estimates than stimuli values. However, given the very poor fit of both models, the argument is not strong. Furthermore, the effe ct is not found when pre-task estimates are used ($z = 0.97$, $p>0.05$).

3.2.3.4. Rank Versus Weighting Function

The central prediction of DbS is that the interpretation or use of an attribute is predicated upon rank order and distribution. Therefore an attribute with a distribution which is not equally represented at all points would result in a non-linear effect upon value estimates. A WADD account would predict that if individuals' use of information is ever non-linear it is due to underlying weighting functions, not the distribution of values. The crime attribute uses a non-linear distribution, while the distance variable uses a non-linear weighting function. To test for these non-linear effects, a multiple regression was performed with both quadratic and cubic components included for all three continuous variables. Separate regression models were calculated for each participant and beta weights for each attribute were tested for significance across participants using a one sample t-test with a null hypothesis of zero. The mean standardised betas and p-values are shown in Table 3.2. There was a significant effect of the cubic component for both crime and distance. In order to plot the effect graphically a further four regression models were computed. Each of these were computed with one of the attributes removed from the model (along with its polynomial terms). By comparing the residuals to those of the original model it is possible to plot the effect of each attribute (Figure 3.4). This shows that although the effect for the crime variable is a close visual match for the distribution of attribute values, rank order (and thus DbS) predicts the reverse curvature (Figure 3.2).

Table 3.2Results of regression analyses when predicting value estimates

	beds	crime	size	Distance	Crime^2	Crime^3	Size^2	Size^3	Dist ^2	Dist ^3
Beta	0.6445	-0.8868	1.9924	-1.2812	1.6965	-0.9990	-3.292	1.5176	1.7476	-0.777
p-value	<0.0001	<0.0001	0.0227	<0.0001	0.0001	0.0001	0.0559	0.0847	0.0003	0.0037

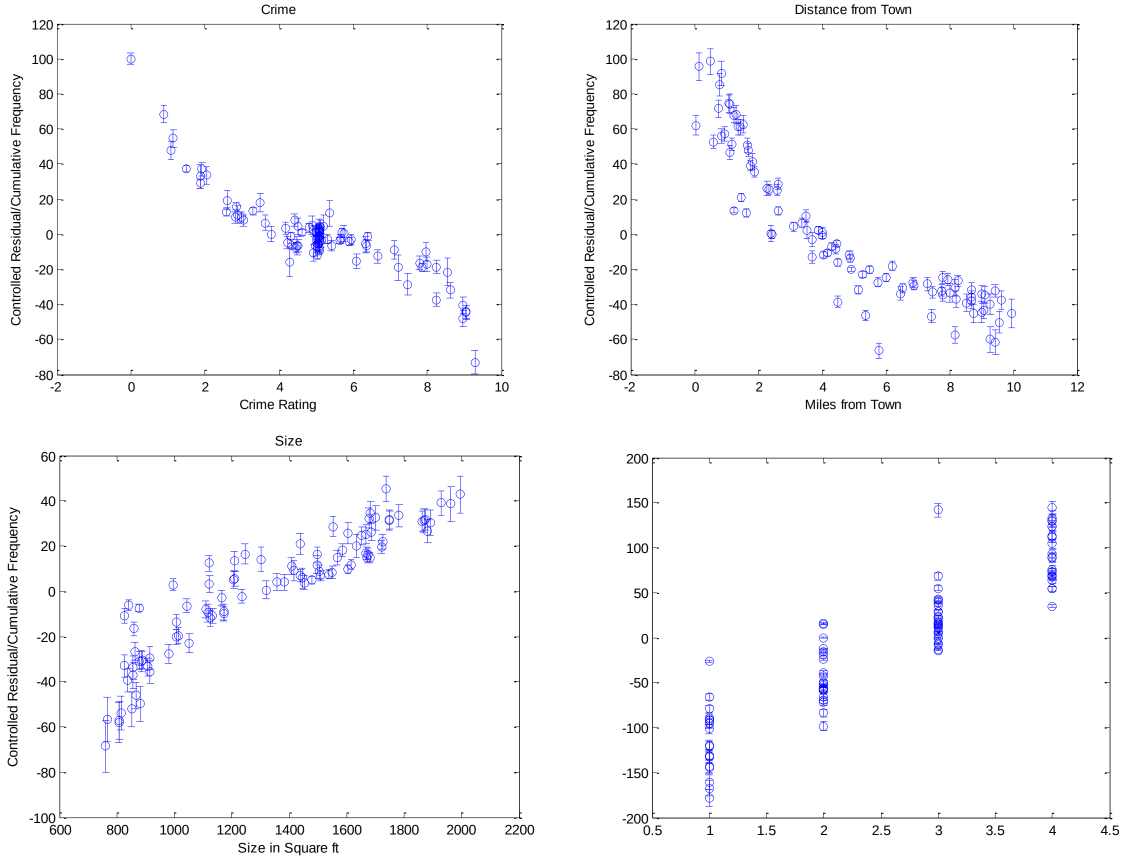

Figure 3.4 Change in residuals caused by each attribute being added to the regression

3.2.4. Discussion

The results show that when calculating value estimates, participants were making good use of the information available and were weighting attributes in accordance with their effect upon true value. However, they were then very poor at estimating these weightings and subjectively reporting the relationship between attributes and item value. Value estimates also show a significant cubic relationship with crime, the attribute whose experienced values had a cubic distribution. However, the curvature of this cubic function was the inverse of that predicted by rank order.

Correlations between revealed attribute weightings and those of the true model show that participants were accurate at weighting attributes and integrating information in accordance with the true underlying model. This suggests that in this multi-attribute valuation task participants were making use of a Weighted Additive model rather than relying solely upon simpler heuristics. A criticism of previous studies using similar estimation methodologies is that they are especially prone to over-fitting the data (Busemeyer & Diederich, 2010; Pachur et al., 2012). In this case the larger numbers of free parameters in the regression model allow it to find marginally better fits than a simpler model, even if the simpler assumptions are true. The estimation procedure does so by fitting a complex model using a large number of cues that then also explains or fits the noise in a specific data set. The analysis employed here reduces these risks by estimating parameters separately for each participant. This finds reliable patterns in parameter estimates across all participants, suggesting that it is not merely an effect of fitting response noise in a single data set. Furthermore, these parameter values correlate significantly with the specific predictions of the WADD model; a pattern unlikely to be the result of simple over-fitting.

Although participants made accurate use of information when making their value estimates, the results show that they were not able to accurately introspect about their own weighting of information. As in previous studies, there was no significant correlation between estimates of attribute importance and the relative weight participants placed upon them during the task (Cook & Stewart, 1975; Reilly & Doherty, 1989; Snook et al., 2011; Zhu & Anderson, 1991). Previous studies of choice have found that when participants can be categorized as using heuristics strategies, they can accurately identify the most important single attribute (or two most important). However, this experiment found that only a relatively small minority of participants are able to identify the attribute they weight most highly in their decisions. For post-task estimates the proportion was actually slightly below chance. The majority of heuristic accounts predict that individuals make use of only a small number of attributes and thus should be able to identify these, as has been shown in other tasks (Snook et al., 2011).

The poor accuracy of participants' explicit estimates of attribute importance means the weighted models of DbS cannot be considered informative. It is not possible to know whether the improved performance in Experiments 1&2 was due to the large number of free parameters. Alternatively it could still be that the underlying model is sound and that the estimation in previous experiments has found true underlying parameter values.

The experiment allowed for a test of DbS's prediction of rank order encoding of attribute scales. The crime attribute used a non-linear distribution, with values in the mid-range of the scale being relatively over-represented. A cubic regression showed that participants' weighting of this scale did show a significant cubic function. However, plotting this effect reveals that the curvature is in fact in the opposite direction to that predicted by DbS. Rank order encoding predicts a steep slope in the mid-section of the scale where the high concentration of items makes the relative rank climb quickly, but a plateau at the extremes where there are few items. What is actually found is a plateau in the centre of the scale and values at the extremes being relatively over-weighted.

One potential explanation for the crime weighting function could be that because the extreme values are comparatively rare, when they are seen they are a more salient cue. Therefore participants adjust their estimate up or down more than they otherwise would. However, a simpler potential explanation is that the observed pattern matches the most reasonable prior expectations. It seems reasonable to assume that the very safest locations will be in locations such as gated communities which are disproportionately expensive, whilst the most dangerous properties will be in especially dilapidated areas and therefore substantially cheaper. This suggestion is also supported by the weighting of an apartment's distance from town. Although the cubic regressor does have a significant effect upon estimates it is significantly weaker than that for crime, with the quadratic component absorbing much of the variance. The revealed pattern is that of apartments particularly close to the town centre being valued highly, but the effect of moving further away declining rapidly. It seems a reasonable assumption that city centre apartments will be disproportionately expensive. The impact of prior expectation and experience is investigated in Experiment 4.

3.3. Experiment Four: Mineral Valuation

3.3.1. Introduction

This experiment used the same stimuli and procedure as in Experiment 3, but used an entirely different cover story. The potential confounds of prior expectations and experiences outside of the experiment environment were eliminated by instead asking participants to estimate the value of mineral deposits at mining sites. In the design of Experiments 1-3, it was assumed that the undergraduate subject pool would have minimal experience of apartment valuation. Furthermore, based upon previous evidence it was predicted that the more recent experiences of the stimuli items were far more likely to be sampled from memory (Ebbinghaus, 1913) and more likely to affect judgements (Beckstead, 2008; Stewart, 2009; Ungemach et al., 2011; Vlaev & Chater, 2007). However, the results of Experiment 3 are consistent with reasonable prior expectations and both Experiments 1&2 find no effect of recency when memory decay functions are added to the DbS model.

The procedure used in this experiment makes it possible to test the cause of the cubic weighting of crime risk represented in valuations during Experiment 3. If this was caused by the relatively rare extreme values being more salient, then the effect will remain. If the effect was simply due to prior-expectations about the value of apartments then the effect will not be present.

It is also possible to check whether participants are still able to accurately weight attributes when they have minimal existing knowledge about their relationship to item value. The task instructions only informed participants of the valence of each attribute's effect, thus participants would have to extract all other information from task feedback. A significant correlation between elicited weighting and true weights would provide evidence that participants are learning from the feedback provided.

3.3.2. Methods

3.3.2.1. Participants

Twenty six undergraduates at the University of Nottingham participated in return for course credit. The average age was 19.5 (SD = 2.4), 9 were male and 17 female.

3.3.2.2. Stimuli

The stimuli used were the same as in the previous experiment but were re-labelled. Instead of the rental value of apartments participants were now told they would be estimating the value per tonne of ore deposits at different mining sites. They were told that the ore contained a fictional mineral called milderite which was crucial to the processing of other precious metals, predominantly platinum. Each potential mining site had various characteristics which would either increase the value of the deposits or decrease them by virtue of increasing the costs of extracting the valuable mineral. Number of bedrooms was re-labelled UN government stability rating, with a better stability rating reducing costs and risks of operating in the region. Floor space became grams of extractible platinum per tonne of ore. Crime and distance from town were replaced with the severity of two contaminants: “gibbsite” and “ferrite” respectively.

3.3.2.3. Procedure

The procedure was exactly the same as in the previous experiment except that now participants were estimating the value per-ton of milderite ore at a series of potential mining sites.

3.3.3. Results

3.3.3.1. Accuracy and Learning Rates

Correlation showed that value estimates were significantly less accurate than in Experiment 3 ($r = 0.599$, $z = 6.62$, $p<0.001$). Plotting average error rates over time reveals that participants initial error rates are greater than in experiment 3 and initial learning appears somewhat slower (Figure

3.5). This is unsurprising given the novel task and lack of valid expectations regarding the unfamiliar mineral deposits.

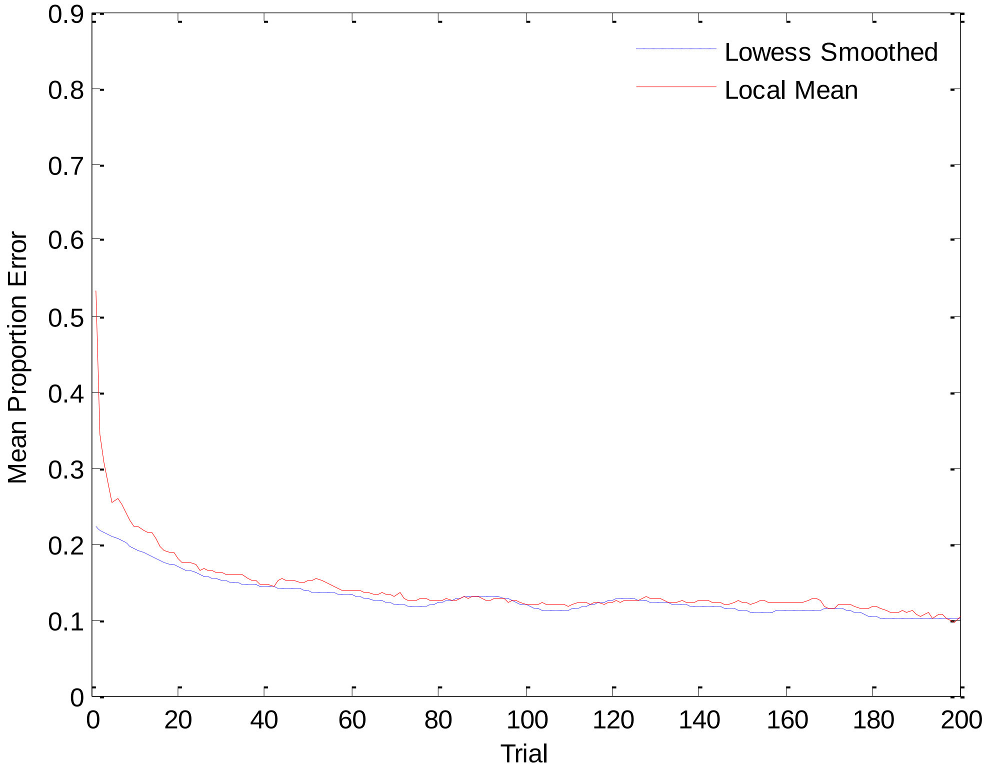

Figure 3.5 The Average error rate over the duration of the Experiment

3.3.3.2. Attribute Weighting Estimates

Standard beta weights were extracted using multiple linear regression. A separate regression model was calculated for each individual. These beta weights correlated strongly with those calculated using the true deposit values ($r = 0.87$, $p<0.001$), but accuracy was worse than in Experiment 3 ($z = 2.06$, $p<0.05$). Unlike Experiment 3 the extractible amount attribute (formerly size) is now weighted accurately (Table 3.3) and is the only attribute that is not significantly underweighted ($t(25) = 0.15$, $p>0.05$). A series of two sample t-tests revealed that there is no significant change in the weighting of gibbsite/crime ($t(47) = -1.03$, $p>0.05$) or ferrite/distance ($t(47) = 1.68$, $p>0.05$). The increase in extractible amount/size is significant ($t(47) = 3.62$, $p<0.001$) as is the increase in the use of stability/bedrooms ($t(47) = 3.9$, $p<0.001$).

Table 3.3 The standardized betas when regressing stimuli values against true values, average estimates accross participants and the mean betas when regressions are performed separately for each participant. These are shown alongside the average subjective estimates of importance for each attribute.

	Truevals	Average Estimate Weightings	Initial estimate	Final estimate
intercept	660.2	623.1		
Stability	0.770	0.460	64.7	74.4
Extractible Amount	0.312	0.316	77.0	66.5
Gibbsite	-0.406	-0.178	69.1	77.9
Ferrite	-0.412	-0.230	67.4	72.7

Standardized betas from true deposit values were correlated (using cluster correction) with individuals' estimates of attribute importance. This revealed no significant correlation with the pre-task estimates (r = -0 008, p>0.05) nor post-task estimates (r = -0.072, p>0.05). There was also no significant relationship between importance estimates and participants own extracted betas for pre task (r = -0.004, p>0.05) or post-task estimates (r = -0.08, p>0.05). The proportion of participants correctly identifying the most influential attribute did show an increase from 23.1% in pre-task estimates to 38.5% post-task. However, the proportion of participants giving their highest rating to the attribute they used least also rose from 11.5% to 34.6%.

3.3.3.3. Decision by Sampling and Modelling

Decision by Sampling's predictions were calculated and correlated with participants' value estimates. This revealed predictive accuracy was not significantly different to that for true values (r = 0.572, z = -1.53, p>0.05). DbS scores were then calculated separately for each attribute and entered into separate multiple regression models for each participant to find the best fitting weighting functions. These were then used to calculate the overall model performance (r = 0.71). However, this was again no different to that found if stimuli values were used instead (r = 0.71).

To reduce the number of free parameters, participants' estimates of attribute importance were used to weight each attribute within the DbS model. When pre-task estimates were used performance was significantly worse than the original DbS model (r = 0.47, z = -4.95, p<0.001) and this was also the case when using post-task estimates (r = 0.53, z = -2.01, p<0.05). Performing the same model using stimuli values reveals a fit which is significantly worse than DbS for pre-experiment estimates (r = 0.41, z = -2.75, p<0.01) and trending in the same direction for post-task estimates (r = 0.50, z = -1.83, p = 0.06).

3.3.3.4. Rank Order Versus Weighting Function

In order to test whether the cubic distribution or weighting functions of stimuli attributes were represented in participants' value estimates, a series of cubic regressions were performed in the same manner as the previous experiment (Table 3.4). This time the attribute with the modified distribution (gibbsite) has no significant cubic effect upon value estimates whereas the attribute with a cubic weighting function (ferrite) does (Figure 3.6). However, the strongest and most statistically significant cubic function is found for the variable for which both distribution and weighting function were linear (extractable amount). Plotting the change in residuals as a result of the attributes being removed from the model reveals that this pattern very closely matches that of crime in Experiment 3 (Figure 3.6).

Table 3.4 Regression results for mineral value estimates

	Stability	Gibbsite	Extractable Amount	Ferrite	Gibbsite ^2	Gibbsite ^3	Extractable Amount ^2	Extractable Amount ^3	Ferrite ^2	Ferrite ^3
Beta	0.46	-0.56	6.32	-0.72	0.78	-0.42	11.56	5.63	1.19	-0.72
p	<0.001	0.005	<0.001	0.003	0.084	0.126	<0.001	<0.001	0.027	0.024

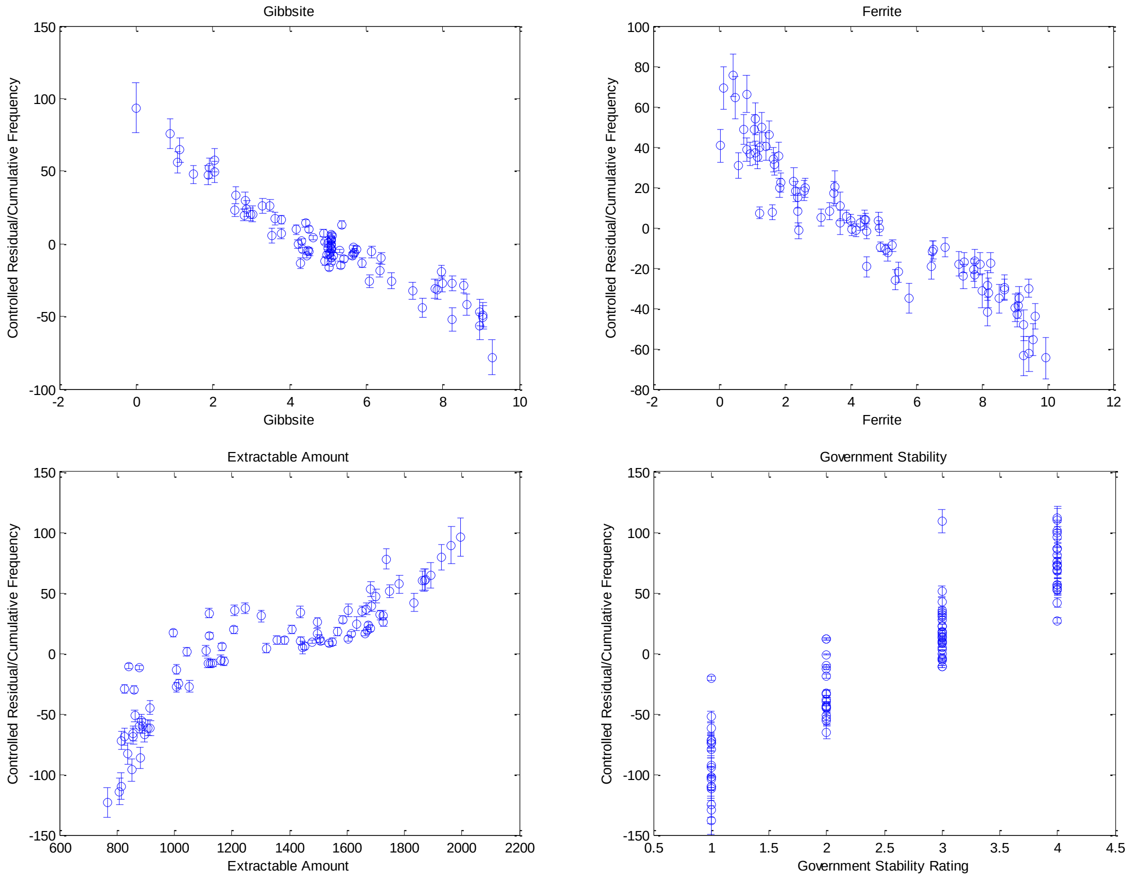

Figure 3.6 Change in residuals caused by each attribute being added to the regression

3.3.4. Discussion

This experiment investigated the extent to which prior expectations and knowledge regarding the stimuli and task shaped the responses and results of Experiment 3. The results show that the change in cover story had a significant effect upon the shape of participants' attribute weighting functions. Gibbsite (formerly crime) had a cubic distribution and therefore DbS predicts a cubic influence upon value estimates. The cubic effect found in Experiment 3 (in the opposite direction to that predicted by DbS), is no longer present after the change in cover story. Ferrite (formerly distance) had a linear distribution and cubic weighting function, so the cubic function found in participants' valuations in both experiments is predicted by a polynomial WADD model. However, no model predicts the strong cubic component found in the current experiment for extractible amount (formerly size) as it had both a linear distribution and linear weighting function. As in Experiment 3, these cubic functions show extreme values being relatively over-weighted in estimates. Therefore a potential explanation is that now participants cannot make use of existing knowledge about the attributes, the extreme values become more salient. These more extreme values can then act as a qualitative cue to raise or lower estimates in a manner not captured by a linear WADD model. As extractable amount and ferrite are both weighted more strongly than gibbsite it seems that this saliency is not equal between attributes, but stronger for those which are more highly weighted. Furthermore the cubic distribution of gibbsite means there are fewer data points at the extremes of the scale, the regions most crucial to the analysis' power.

The finding most consistent with Experiment 3 was that participants were poor at estimating their usage and weighting of attributes. Furthermore, the accuracy with which participant valuation strategies matched the underlying relationships between attributes and true value was still high. Accuracy was significantly lower than the previous experiment, so prior expectations certainly had an influence. Although this influence appears to be relatively minor, the fact that accuracy over time does still asymptote before the mid-point of the task, suggests it may not simply be an effect of slower learning when there are no suitable pre-existing hypotheses (Harvey & Fischer, 2005).

3.4. Chapter Discussion

This chapter used multi-attribute valuation tasks to investigate how individuals make use of the relative rank of attribute values and the underlying mathematical relationships linking attributes to overall item value. It also examined the degree to which responses are shaped by expectations of real world values and experiences which occur prior to the task. This was achieved by changing the cover story, so although participants believed they were valuing a very different class of items the stimuli and values were in fact identical in each case.

3.4.1. Non-Linearity in Weighting and Distribution Functions

In Experiment 3 participants estimated the rental value of flats. The distribution of the crime risk attribute was shaped such that rank order encoding would predict a cubic curve, with an increase in value in the mid-portion of the scale providing a disproportionately large increase in DbS score compared to at the scale's extremes. Although a cubic function was found in participants responses, the curvature was in the wrong direction. There was a plateau in the mid-portion of the scale, with an increase in attribute value providing a disproportionately small increase in estimates of item value. Two suggestions seemed to be reasonable hypotheses: One was that the extreme values on the scale were made more salient by their relative rarity and were thus over-weighted when they were seen. The second was that participants could simply be using expectations from experience in the real world. This was because it seems reasonable to assume that the very safest and the most dangerous locations are disproportionately expensive and cheap. The latter was the favoured hypothesis, as the weighting of distance from town had a particularly strong quadratic curve and matched reasonable real world expectations: Locations very close to town were valued disproportionately highly, whilst the effect of moving further away from town diminished rapidly. The same pattern as city centre flats being disproportionately expensive.

In Experiment 4 the same stimuli were used but participants were instead told they were estimating the value of deposits of a fictional mineral "milderite". This largely eliminates the effect of prior expectations. Therefore if both experiments produced the same cubic encoding effects it could only be attributable to the distribution of attribute values. However, participants weighting of gibbsite (formerly crime), the attribute with a cubic distribution, became linear. Previous studies using multi-attribute tasks have shown that prior task knowledge can have a significant effect upon learning and accuracy (Balzer et al., 1989; Harvey & Fischer, 2005), so in this light these results are unsurprising. However, it was predicted that the effect of prior experience would be minimal and largely overshadowed by the values experienced in the first 100 trials (Beckstead, 2008; Ebbinghaus, 1913; Stewart, 2009; Ungemach et al., 2011; Vlaev & Chater, 2007).The observed effects of prior expectations contradict the prediction that more recent experiences have a disproportionate effect upon judgement. This is also true of the modelling performed in experiments 1& 2 which showed no improvement when memory phenomena were incorporated into sampling models. Therefore, if judgements are made based upon comparisons with previous experiences it seems that more recent experiences do not have the disproportionately large effect predicted by much of the memory literature.

Although the pattern of encoding for distance from town supports prior expectations as an explanation, it cannot explain the cubic encoding of other scales in Experiment 4. The most significant cubic component (both statistically and in terms of size) was found for the linearly distributed and linearly weighted attribute: extractible amount. This was also the only continuously distributed attribute whose weighting in participants judgements significantly increased from Experiment 3. It seems an alternative explanation is that when participants saw an extreme value for an attribute they believed was influential, this would act as a cue to raise or lower their value estimate in addition to any WADD calculations. Gibbsite is the only continuous variable which does not display this pattern and it is also the one which has the lowest overall weighting in participants' responses. Its distribution also means that the extremes of the scale are relatively under-represented in the data and this serves to reduce the sensitivity of the cubic regression. The same argument cannot be made for Experiment 3 as crime also receives low weighting, but in this case a strong cubic component, hence the positing of different explanations for each experiment.

3.4.2. Attribute weighting estimates and accuracy

In Experiments 3&4, participants' value estimates showed they were making appropriate use of information. The weight given to each attribute closely matched the relationships between attributes and true values suggesting that participants were using the feedback to accurately abstract the underlying model. This is a pattern which supports WADD and cannot be easily explained by heuristic accounts without relying upon criticisms such as over-fitting (Gigerenzer & Todd, 2000), the risks of which are minimised in these experiments.

Despite making accurate use of the stimuli information when calculating their valuations, participants were extremely poor at providing subjective estimates of their own attribute weightings. This was true both before and after completing the task. This suggests that the cognitive process is not available to introspection. This is not surprising in itself, given the existing literature (Cook & Stewart, 1975; Reilly & Doherty, 1989; Snook et al., 2011; Zhu & Anderson, 1991) and that introspection is not explicitly predicted by rational models. What is somewhat surprising is that participants were very poor at simply identifying the attribute they weighted most highly. Previous research has suggested that when participants are employing a heuristic strategy they are able to identify the one (or two) attribute(s) upon which they base their responses (Dhami, 2003; Snook et al., 2011). However, participants were not explicitly asked to identify the most important, this was inferred from their estimates of attribute importance provided on a continuous scale. It is possible that using an elicitation method more compatible with heuristic strategies will enable participants to improve their introspection, but this seems unlikely given the lack of other evidence for heuristic strategies in this task.

3.4.3. Implications for Decision by Sampling

Overall the results of this chapter do not support DbS. The use of attribute scales did not conform to rank order encoding. In Experiment 3 there was cubic encoding of the predicted attribute scale, but the curvature was in the opposite direction to that predicted by the model. In Experiment 4 there was no cubic encoding of the predicted scale and there was cubic encoding of attributes where it was not predicted; though the curvature was again in the wrong direction for the model. As in Experiments 1&2, modelling DbS revealed poorer accuracy than a simple baseline measure and when attribute weights were allowed to vary as free parameters the model performed no better than WADD. The results of setting these free parameters using participants' explicit estimates was somewhat uninformative as the results reveal participants are unable to accurately introspect about these weightings.

DbS is fundamentally a model of choice rather than valuation. The results presented in this chapter and the preceding one suggest that DbS cannot accurately predict value estimates in a multi-attribute task. However, there is a large body of evidence suggesting that there are qualitative differences between strategies of single item valuation and choices between alternatives (Hsee, Loewenstein, Blount, & Bazerman, 1999; Lichtenstein & Slovic, 1971; Tunney, 2006; Tversky, Slovic, & Kahneman, 1990). In the next chapter participants are asked to make a binary choice between two alternatives and the accuracy of DbS is tested when applied to predicting decisions.

4. Chapter 4

4.1. Chapter Introduction

Decision by Sampling is fundamentally a model of choice and was not specifically designed to explain single item valuations. Therefore the poor performance of this model in previous chapters could be because individuals use qualitatively different strategies for the different tasks (Gigerenzer & Selten, 2002; Lichtenstein & Slovic, 1971). This chapter presents two experiments which investigate decisions in a multi-attribute choice task.

Experiment 5 compared the performance of a mathematically compensatory model of choice with a simple heuristic rule. It also tested the performance of DbS, investigating whether it can provide an appropriate link between mathematical and heuristic models. It tested the hypothesis that memory effects and sample size can explain why individuals are frequently categorized as using one of two very different cognitive strategies. The strategies for valuation and choice tasks were also compared, testing two different accounts of the differences between them: An adaptive toolbox of heuristics (Gigerenzer & Selten, 2002) and a single cognitive system with differences caused by the interpretability of the information in different tasks.

Experiment 6 used eye tracking to examine the accuracy of drift diffusion models of choice which make specific predictions about individual's patterns of attention to information and the resulting effect upon choices. The behavioural effects of skewed distributions upon choice were also examined in order to test the predictions of DbS. Participants' use of information was modelled to examine whether they made use of an attribute's weighting function or its distribution.

4.2. Experiment 5: Binary Comparisons and Weighting Functions

4.2.1. Introduction

4.2.1.1. Compensatory Calculations and Heuristics

Traditionally, models of value based decision making have fallen into one of two categories: Complex mathematically compensatory models and simple heuristic rules. The former includes models in which the decision maker calculates a score or utility for individual items using a system of weighted combination and trade-offs across all (or at least a large proportion) of the cues presented (e.g. LENS model; Brunswick, 1955; Weighted Additive rule/Franklin's rule, Gigerenzer & Todd, 2000; Cumulative Prospect Theory; Tversky & Kahneman, 1992; and Expected Utility Theory; von Neumann & Morgenstern, 1944). The models are described as compensatory because if an item has a low value on one attribute it can be compensated for by a high value on another. Heuristic accounts posit that decision makers use simple rules and that there is no (or minimal) mathematical computation (e.g. Priority Heuristic; Brandstätter et al., 2006; Matching Heuristic; Dhami, 2003; Take The Best; Gigerenzer & Goldstein, 1996; and Elimination by Aspects; Tversky, 1972). A popular interpretation of heuristic accounts is the adaptive toolbox (Gigerenzer & Selten, 2002). This suggests that there are a potentially large number of simple heuristics which decision makers have at their disposal and that the task environment dictates which is used in a deterministic bottom up manner. Therefore, when the decision environment or the response mode changes, individuals will use different heuristics in the same manner that a camper will use a different tool on a Swiss army knife depending upon the problem they are faced with.

Despite initially promising findings for some models of heuristic decision making (Brandstätter et al., 2006; Dhami, 2003; Gigerenzer & Goldstein, 1996), recent studies have found that these models perform poorly when items are created so that compensatory and non-compensatory models make opposing predictions (Birnbaum & LaCroix, 2008; Glöckner & Betsch, 2008; Johnson et al., 2008; Rettinger & Hastie, 2001). Although the weight of evidence would seem to support compensatory models there are nonetheless a number of findings that these models alone cannot explain. When subjects are categorized according to the models that best explain their decision making (Bröder, 2002) the majority of individuals are identified as using a compensatory strategy. However, a significant minority are best explained by a non-compensatory heuristic or rule (Ayal & Hochman, 2009; Bröder, 2003; Glöckner & Betsch, 2008; Glöckner & Herbold, 2011; Newell & Shanks, 2003). Of course there remains the possibility that rather than falling into a distinct taxonomy, subjects may be using a dual process strategy that incorporates both compensatory valuation and a set of heuristic rules (Ayal & Hochman, 2009). Or individuals could actually be employing a different strategy altogether, one which lies somewhere in between the two extremes.

The experiments reported here investigated mathematical and heuristic strategies using the Weighted ADDitive rule (Gigerenzer & Todd, 2000) and Dawes' Rule (Dawes, 1979). Despite sometimes being referred to as a rule, WADD is actually a mathematically compensatory calculation whereby weights are applied to each attribute according to their relative importance. The values of each attribute are then multiplied by their respective weighting function and summed together to enable a single comparison between the options. Dawes' rule states that decisions are made by assessing each attribute in turn and making a binary judgement of which item has a higher value for that attribute. Whichever item is better has its score incremented by 1. If the attribute does not discriminate between the items then neither score is incremented. Once all the attributes have been assessed the item with the highest score is chosen. The overall Dawes' score for each choice can also be calculated by subtracting one item's score from the other.

Although Dawes' Rule was originally conceived as a prescriptive theory of how to improve decision making in complex real world situations, some descriptive theories of choice imply that there are situations where Dawes' is a useful description of choice. One case in particular is Decision by Sampling (Stewart et al., 2006). Dawes is actually a special case of DbS, where the sample is so small that only the currently available information is used to calculate an item's score. Therefore, a valid question to ask is whether individuals should not be described as using qualitatively different strategies, but are in fact using DbS but calculate scores using different sample sizes. This is potentially a more parsimonious explanation as a single model could apply to all individuals. Only a single parameter would need to vary in order to explain individual differences.

4.2.1.2. Valuation vs. Choices

A question inherent in the mathematical versus heuristic debate is whether value estimates for a single item and choices between two or more alternatives are made using the same cognitive process. Mathematical models generally assume that the same cognitive process can be used to make value judgements and to make choices (Brunswick, 1955). They suggest that any apparent differences between choice and value judgements occur because of the interpretability of information being inherently different in the two tasks i.e. that the inputs are different, rather than the process itself. An example would be selecting a new set of speakers. When in a shop and comparing the sound of several different models it is easy to judge the relative sound quality of each model. However, if one is trying to assess the sound quality of a single set in isolation, it is very difficult to quantify how good it is with nothing to compare it against (Hsee et al., 1999). When then selecting the preferred option or judging the value of a single set, the individual is still weighing up the evidence in the same manner, but the evidence which has been accumulated is respectively more or less accurate. Conversely, a heuristic or adaptive toolbox account would posit that these differences arise because any one heuristic is specific to the task environment and there is no manner in which the same heuristic or decision rule can be applied to both response types.

A strong line of argument in favour of a qualitative difference is the phenomena of preference reversals. This is where individuals express a choice preference for the item to which they previously assigned a lower value when they saw it alone (Lichtenstein & Slovic, 1971; Tversky et al., 1990). These findings have proven very robust and extend across a wide range of decision domains (Hsee et al., 1999; Tunney, 2006). However, most bid-choice reversals are found to occur when at least one attribute value is difficult to evaluate in isolation, with relative differences only becoming apparent when an item is compared alongside others in the choice condition. This suggests that preference reversals may not require an explanation based upon qualitatively different processes. Studies that have assessed participants' intensity estimates of single scales or attributes find that these individual scales are interpreted differently when presented alone and when presented with other comparable values. This difference in interpreting or encoding individual attributes then accounts for subsequent inconsistencies between choices and valuation of the multi-attribute item(s), (Johnson, Haubl, & Keinan, 2007; Sevdalis & Harvey, 2006). Either account allows for the possibility that DbS will provide a good explanation of decisions, despite experiments in previous chapters demonstrating poor predictive ability on valuation tasks.

This experiment employed a multi-attribute choice task where subjects both value the individual items and then also make binary choices between pairs of items. Crucially, the pairs are constructed so that the predictions of the Dawes' Rule and mathematical calculations of objective value are in opposition for a significant proportion of choices. We also compare the effect of attribute values upon both choices and valuation. If qualitatively different processes are recruited then one would expect only a minority of attributes to be used during choice and either a large majority or a different selection of attributes to influence single item valuation. However, if the same process is recruited to both judgements, then any differences will be in the manner in which all attributes are weighted. Two attributes are manipulated such that we can examine the relative effects of mathematical relationship with an item's overall value and the distribution of experienced attribute values. A standard WADD account predicts the former will be most influential, whilst DbS predicts it will be distribution.

4.2.2. Methods

4.2.2.1. Participants

Thirty-two students from the University of Nottingham participated in the study, 7 were male and 25 female. The average age was 22 years (S.D. = 4.9). Participants were paid £8 for participation.

4.2.2.2. Stimuli

Stimuli consisted of 125 hypothetical apartments and from these 124 pairs were selected based upon criteria detailed below. Each stimulus item consisted of five attributes that plausibly influence the rental value of apartments: the number of bedrooms, the number of bathrooms, the floor size in square feet, the distance from the town centre and the crime risk.

Apartments were again used in this experiment because chapter 3 demonstrated that participants' accuracy in estimating target value was significantly higher than for more abstract stimuli such as mineral deposits. Furthermore, all of the relevant hypotheses which this experiment tests are either orthogonal to the effects of prior knowledge or make predictions which are in the opposite direction to those demonstrated in Chapter 4.

The number of bedrooms was a random number between 1 and 4 and the number of bathrooms was a random number between 1 and the number of bedrooms. The floor size was randomly selected between 750 and 2000 square feet and distance from town was also random, between 0 and 10 miles. Crime risk was given as a score between 0-10 and explained to subjects as the crime rating of the property's postcode as calculated by the government statistics website "police.uk". The distribution of crime risk values was derived using distance from town: the distances given to all stimuli were de-meaned, such that the distribution was centred at zero. These values were then cubed before the being re-scaled to between 0 and 10. This created a non-linear distribution with a cubic curve, meaning that values in the centre of the range were proportionally over-represented whereas the extremes were under-represented (Figure 4.1).

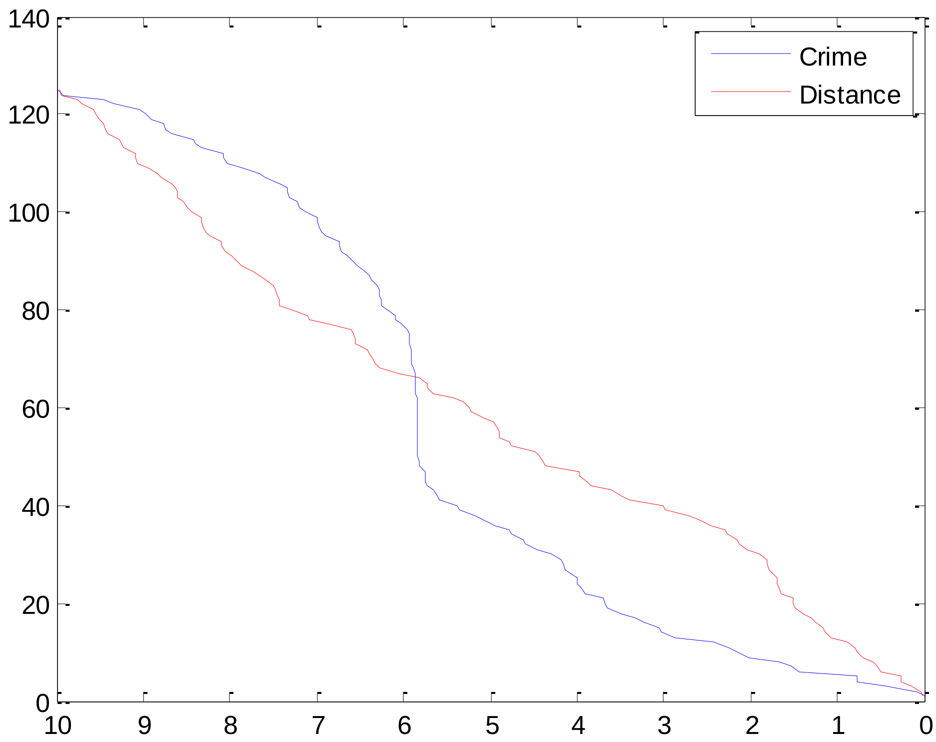

Figure 4.1 The cumulative frequency plots for the two modified attribute scales

The rental value of each stimulus item was calculated from a base of £400. Each attribute had its own weighting function used to calculate its effect upon rental value, as shown below. The output of each was then multiplied against the starting value of £400 to obtain the rental value for each apartment.

Equation 10

$$1 + \frac{(beds - 1)}{4} * 0.15$$

Equation 11

$$1 + \frac{(baths - 1)}{4} * 0.125$$

Equation 12

$$1 + \frac{(size - 750)}{1250} * 0.2$$

Equation 13

$$1 + \frac{(10 - crime)}{5} * 0.5$$

The cubic function used to create the non-linear distribution of crime risk was then used to create the weighting function for distance. The result is that both distance and crime have the same overall effect upon rental values, but the cubic and linear components are provided by either distribution or value weighting function respectively.

The stimulus pairs were then selected from the resulting 125 items. The main aim of this experiment was to investigate choice behaviour when the score derived from a direct binary comparison of each attribute (hereon "Dawes score") favoured the item with the lower objective rental value. We will refer to these as "mismatch trials". All potential pairings of all 125 stimuli were analysed as mismatch trials using the following criteria: Pairs where there was a mismatch and where the Dawes score was as high as 3, were required to have a difference in absolute value of at least £50. These were inevitably rare and only 4 pairs met the criteria. For mismatch pairs with a Dawes score of 2 and 1, we required a minimum value difference of £100 and £200 respectively. We therefore selected 46 pairs with a Dawes score of 2 and 27 with a score of 1. The remaining 47 were "matched trials", where Dawes score favoured the item which also had the higher objective rental value. These pairs were selected entirely at random, the only criteria was that there were no duplicate pairs within the stimuli set.

4.2.2.3. Procedure

The experiment took the form of a simulated estate agent (realtor) software package for apartments in the rental market. The subjects were told that the apartments were situated in a typical British city of a similar size and standard of living to the one in which the experiment was conducted (Nottingham), but that there were some differences which they would learn as they progressed. In the first part of the experiment the subjects saw each of the 125 stimulus items individually and were asked to estimate the monthly rental value of each one. The five attributes of a single apartment were presented in a list and subjects simply typed their estimate of monthly rent. After each estimate the subjects received feedback in the form of the objective rental value (determined by the formulae described above). The trials were self-paced but feedback was visible for 2000msecs , followed by a 1500msec ITI. In the second half of the experiment the two items in an apartment pair were presented side by side and subjects were asked to indicate which of the two they thought was the more valuable using an on-screen button. No feedback was given during this part of the experiment.

4.2.3. Results

4.2.3.1. Behavioural data

Responses from the first part of the experiment show that subjects were relatively accurate at estimating rental value. Estimates were an average of 14.25% off the true rental values across all individual responses and when estimates were averaged across subjects there was a very strong correlation with true values ($r = 0.947$, $p<0.001$). Estimates also became more accurate over time: The correlation for the first 50 items was weaker ($r = 0.926$, $p<0.001$) than for the remaining 75 ($r = 0.96$, $p<0.001$) and correlating average error with trial number reveals a significant relationship ($r = -0.2$, $p<0.05$).

4.2.3.2. Categorizing Individuals' Behaviour

The next question was which of the candidate models provided the best fit of the observed data. This allows us to see if we have replicated previous reports in which the majority of subjects' behaviour is best explained by mathematical models and only a small minority by heuristic models (Bröder, 2002; Glöckner & Betsch, 2008; Glöckner & Herbold, 2011). Logistic regressions were used to estimate WADD weights for each subject. These weights were then used to predict their choice behaviour. Across all the stimulus pairs the model predicted 88.4% (SD = 7.5) of responses. However, when this analysis was separated by trial type there was a significant difference in the model's accuracy between matched (M = 93.5%, SD = 6.4%) and mismatched (M = 85.4%, SD = 9.5%) trials (t(31) = 5.56, p<0.001). Using the same simulation method Dawes' Rule provided a poorer fit and predicted an average of just 53.5% of choices (SD = 14%) across all trials. The rule also showed good accuracy on matched trials (M = 89.2%, SD = 6%), but a poorer fit for mismatched trials (M = 31.8%, SD = 22%). This would appear to support the mathematical model, as Dawes performs well when it's predictions are in line with WADD, but performs significantly below chance when the two models make opposing predictions. However, this cannot be a complete explanation since that would result in an exact reversal in accuracy between matched and mismatched trials i.e. 11.8% in mismatched (100-89.2) or 69.2% in matched (100-31.8).

Decision by Sampling was explicitly modelled from the data. For each choice, the number of favourable comparisons within all previously seen choices and single item valuations was calculated for each attribute. These were then summed to create a single score for each item. This resulted in the correct prediction of 78.0% (SD = 10.6) of choices across all trials. For mismatched trials this was 69.7% (SD = 14.5) and matched trials 91.3% (SD = 7.5). The effect of recency was investigated by adding a weighting curve to the DbS model. Thus, a favourable comparison to a more recently experienced item would increment the item's score more than a favourable comparison to a more temporally distant item. The size of this difference was controlled by a weighting function which took the shape of an Ebbinghaus forgetting curve. Maximum likelihood estimation was used to find the most appropriate rate of decay for each participant. This found that the best fitting decay rate on average was 0.032 (SD = 0.0057). This actually produces a concave weighting function, with the vast majority of items weighted very highly and similarly, except for the very oldest. This results in predictive accuracy which is identical to the unweighted model. Closer inspection reveals that the MLE search function terminates once the weighting function becomes so linear that there is no longer any predictive difference between the weighted and unweighted predictions. This is the case for all participants; there is no evidence that different individuals are simply using different sample sizes. As a result, the performance of WADD and DbS are significantly correlated (r = 0.66, p<0.001), but with DbS always performing slightly worse.

As with earlier studies (Bröder, 2002; Glöckner & Betsch, 2008; Glöckner & Herbold, 2011) participants were next categorized according to which model best predicted their responses. Thirty-one (96.9%) of the subjects were best described by WADD. None were categorized as responding according to DBS or Dawes' rule and only one (3.1%) subject was predicted equally well by WADD and Dawes rule. We then analysed the Match and Mismatch trials separately. For the Matched trials 26 (81.3%) subjects were best fit by WADD, 3 (9.4%) by Dawes' and 3 (9.4%) were fit equally well by both, none were fit best by DbS. When only Mismatched trials were examined 30 (94.8%) subjects were best explained by WADD, none by DbS or Dawes' rule and 2 (6.2%) equally well by WADD and Dawes. This pattern of individual differences in decision rules is similar to that reported in other studies that also show that the majority of subjects appear to utilize mathematical compensatory models when making choices between items.

When DbS was modelled with additional free parameters which allowed attributes to receive different weightings there was a marked improvement in accuracy, correctly predicting 86.3% of choices overall. However the model only performed better than WADD for 6 participants and was still significantly worse across participants ($t(31) = 5.09$, $p<0.001$). When split by trial type, DbS was less accurate for mismatched pairs, predicting 84.1%. It was the best performing model for only 3 participants and significantly worse than WADD overall ($t(31) = 4.92$, $p<0.001$). The DbS model allowing different weightings performs better for matched trials, predicting 89.8% of choices and is the best performing model for 16 participants. However, its accuracy is not significantly different from that of WADD, with the relative differences between them forming an approximately normal distribution around zero (mean difference = 0.012, $t(31) = 1.04$, $p>0.05$).

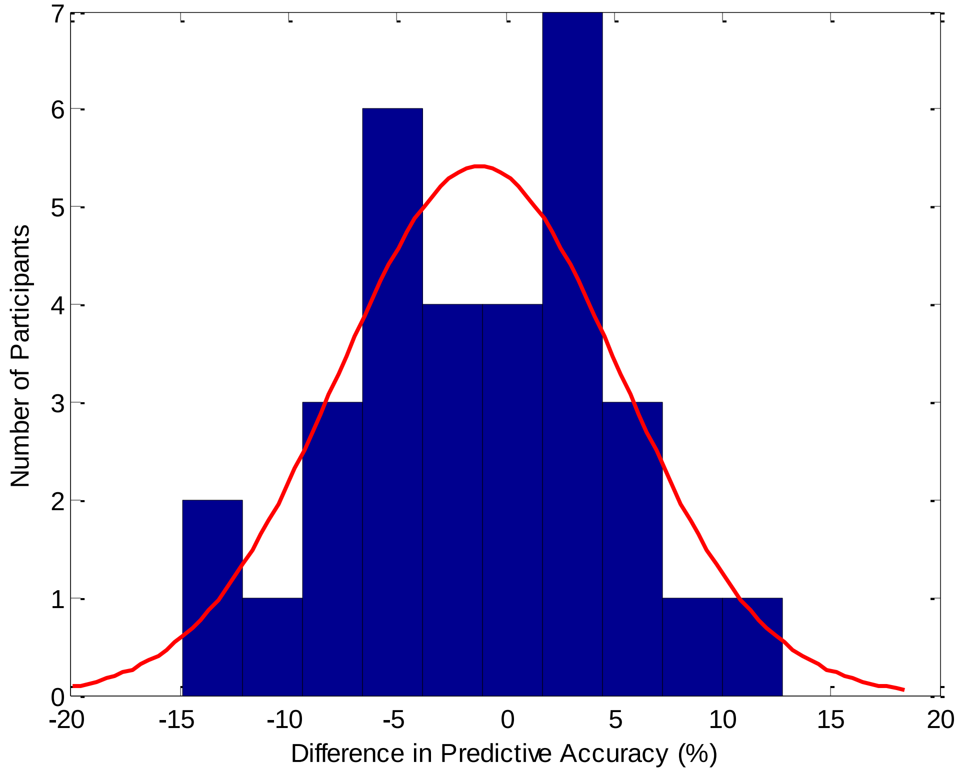

Figure 4.2 Distribution of differences between percentage accuracy for weighted DbS and WADD models. The line shows the best fitting normal distribution curve.

4.2.3.3. Are valuations and choices made using the same process?

Now we investigate whether subjects make similar use of information when providing single item valuations and when choosing between alternatives. If qualitatively different rules or heuristics are used then subjects would use the information differently. If the same compensatory process is used in both tasks then there should only be minimal differences in weightings. To allow for comparison between the continuous data of value estimates and binary choice responses the average choice proportions for each stimulus pair were entered into a linear regression with the difference between items on each attribute entered as predictors. This reveals that the overall pattern of information weighting is very similar in both tasks, with one major exception: The distance to town centre is significantly over-weighted during the choice task. This is in comparison to both the valuation task and to its objective influence upon rental value. This is particularly interesting because this attribute had a curvilinear weighting function with respect to rental value.

Therefore, it suggests that participants were correctly representing this weighting function in the valuation task, but then the constant distribution of values enlarged the perceived difference between the items in the choice task (Table 4.1 & Figure 4.3).

Table 4.1 WADD weights for choice proportions as estimated using a standardized regression

Attribute	Standardized Betas	t-value	p-value
Beds	.668	-13.339	<.001
Baths	.208	-3.901	<.001
Size	.079	-1.549	.124
Crime	-.232	2.588	.011
Distance to Town	-.445	4.899	<.001

Table 4.2 WADD weights for estimates of rental values as estimated using a cluster corrected standardized regression

Attribute	Standardized Betas	t-value	p-value
Beds	.506	17.35	<.001
Baths	.173	8.05	<.001
Size	.084	5.02	<.001
Crime	-.296	12.38	<.001
Distance to Town	-.074	2.94	.006

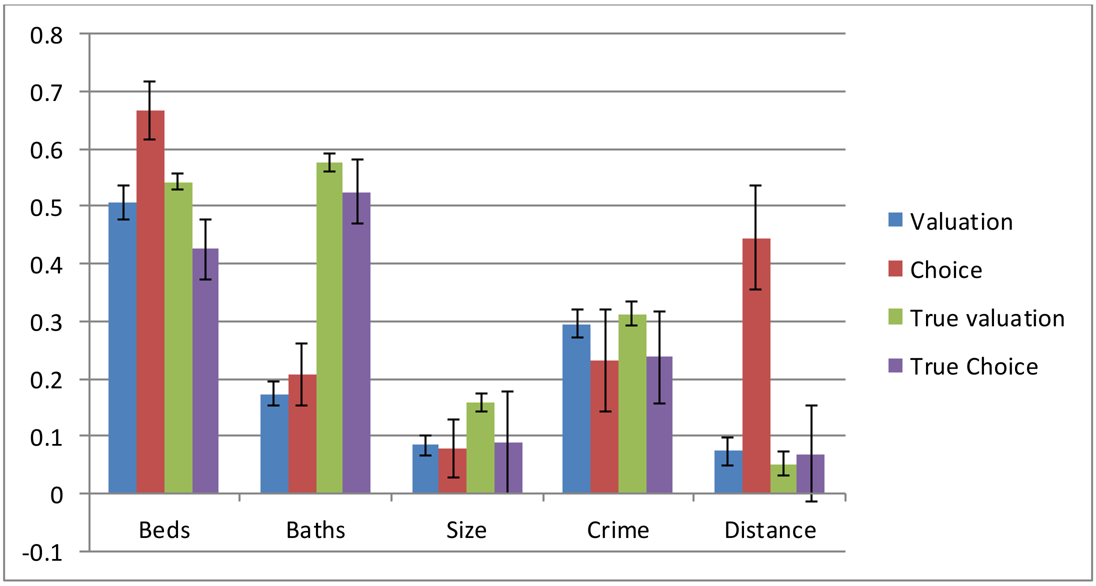

Figure 4.3 WADD weights for each attribute as calculated for choices and item valuation. Error bars represent Parameter Estimates

The experiment had been designed such that participants weighting of crime and distance could be examined along the range of their scales. This would allow a direct test for a cubic component in individuals' weighting of the information and for the resulting pattern to be plotted graphically. However, due to a programming error during stimuli creation there was a highly significant correlation between crime and distance ($r = 0.9$). Therefore, the effect of the two scales could not be sufficiently separated by non-linear modelling as originally planned. Experiment 6 below corrects this error.

4.2.4. Discussion

This experiment examined whether compensatory models provide better explanations of choices than non-compensatory heuristic models and whether DbS can provide a link between the two. The results replicate the basic findings of other studies (Bröder, 2002; Glöckner & Betsch, 2008; Glöckner & Herbold, 2011) which found that the behaviour of a large majority of individuals is best explained by mathematically compensatory models. A very small number were better explained by Dawes rule, but none by unweighted DbS. Accuracy was improved when attribute weightings were allowed to vary within the DbS model but its accuracy was still significantly worse than WADD and it was only the best fitting model for 6 out of the 32 participants. However, within this modelli ng it should be noted that the matched trials provide the fairest comparison between DbS and WADD. In order to provide an accurate comparison between WADD and Dawes rule, stimuli pairs were carefully controlled using the mathematically calculated rental values. Thus, WADD predicted the same average difference between apartment valuations for matched and mismatched trials, i.e. a 1:1 ratio. Because DbS scores are calculated on different information, the ratio of differences in its predicted values was 1:1.5, with the greater difference being found for mismatched trials. Therefore, DbS is at a relative disadvantage and one should be very careful when comparing its accuracy to that of WADD on mismatched trials.

It was hypothesised that DbS could explain why individuals are so frequently split between mathematical and heuristic models. DbS could potentially fit the responses of participants who are otherwise best explained by WADD and Dawes by fitting different size samples. A larger DbS sample would result in high correlation with WADD predictions, whilst a smaller sample would correlate with Dawes. However, as in valuation tasks, the best fitting DbS parameters were those which resulted in the highest correlation with WADD. This was true for all participants, even those for which Dawes rule predicted a high proportion of their responses. There was no evidence for any difference in sampling rates between individuals.

4.2.4.1. Valuation vs. Choice

These results show that subjects made different use of the same information when making single item valuations and when choosing between alternatives. Although the results do not suggest a broad and all-encompassing switch in strategy such as from a weighted compensatory judgement to a simple heuristic based mechanism (Brandstätter et al., 2006; Gigerenzer & Selten, 2002), the difference in the way particular attributes are used is interesting. The largest change in weightings was found for the distance from town attribute. This was the lowest weighted during valuation but second highest in the choice task. Crucially, this is one of the attributes where the relative influence of the experienced distribution and its weighting upon rental value were manipulated. Whilst it was equally represented at all points between 0 and 10 miles, the weighting function relating it to rental value was curvilinear. This means that for the majority of the centre portion of the scale a difference in distance corresponded to relatively little change in value and individuals seem to have represented this in their value estimates. However, the fact that it was equally represented at all points on the scale means that when two items are put side by side there is a greater chance of a large relative difference than there is for other attributes, particularly crime. Thus, the observation that subjects placed far more weight upon it in the choice task suggests that the relative difference is more evaluable or simply more salient during the choice task (Hsee et al., 1999). Unfortunately the programming error in stimuli creation means it is not possible to investigate whether a cubic weighting function can be found in participants use of either scale. This is now addressed, along with other questions, in Experiment 6.

4.3. Experiment 6: Eye Tracking

4.3.1. Introduction

Whether a choice requires selecting which apartment to rent in a new town or just a preferred snack in a shop, these decisions are not made immediately upon exposure to the options. Rather they recruit a form of deliberative decision making not explicitly captured by WADD. This experiment examines a family of models explicitly designed to describe such choice processes: Accumulator models. These generally assume that decision making is based upon a noisy stochastic process where evidence for each item is accumulated over time. A choice is made once enough evidence has been accumulated for one item to cross a threshold (Busemeyer & Townsend, 1993; Ratcliff & McKoon, 2007; Stewart et al., 2006; Usher & McClelland, 2001).

In this experiment we focus upon one class of accumulator model: the drift diffusion model (DDM). DDM predicts that when making a choice between two alternatives individuals rely upon a single accumulator which accumulates evidence over time from each of the competing items (Ratcliff & McKoon, 2007). The value of the accumulator at any given time point represents the current difference in evidence accumulated for either option. A positive value indicates a current preference for item 1, whereas a negative indicates the same for item 2. A decision is made when the accumulator crosses a decision boundary at either 1 or -1.

Although there are many versions and iterations of DDMs (Bogacz et al., 2006), the one which we will focus on here is that described in Krajbich, Armel, & Rangel (2010). This model posits that the rate of drift is a function of attention and relative item value as described by $(V_A - \theta V_B)$ when attention is directed towards item A. This applies the weighting function θ to the value of the unattended item. This weighting can vary between 0 and 1. Therefore, when θ = 1 there is no effect of attention and the rate of drift depends solely upon the options' relative values. However, when θ = 0 then the value of the unattended item is irrelevant and evidence is accumulated for the attended item regardless of the relative appeal of the alternative. Many other accumulator models have a similar assumption, whereby the slope is a function of the information and item currently attended to (Busemeyer & Townsend, 1993). Eye tracking analysis is a useful tool when investigating such models because visual fixations can be used as a proxy for the attention term in the models. We chose to investigate the model above because it has previously performed well when applied to eye tracking data and can be easily extended to multiple attribute choices.

By investigating multiple attribute choices we can extend previous findings and address additional questions of how attention is directed when collecting information and making a decision. One of the most reliable findings in eye tracking research is the gaze cascade. This is the finding that in the lead up to making their selection, individuals look more and more towards the item they subsequently select (Atalay, Bodur, & Rasolofoarison, 2012; Glaholt & Reingold, 2009; Shimojo, Simion, Shimojo, & Scheier, 2003). Several papers argue that this is evidence of a feedback loop between value judgement and attention or saccade planning processes (Shimojo et al., 2003; Simion & Shimojo, 2007). However, it seems desirable that attention is not only directed towards the most valuable option but also towards the more important or useful information. In a multi-attribute choice the relative importance of different pieces of information can differ from one choice to the next. For example, when choosing between a one-bedroom apartment and an alternative with four, one might expect individuals to attend more to this information than on trials where both options have the same number of bedrooms. Thus the multi attribute design means that the effects of item value and importance of information can be separated and the information upon which the attention feedback loop is predicated can be better elucidated.

4.3.2. Method

4.3.2.1. Participants

Twenty-four students at the University of Nottingham participated in return for course credit. There were 17 females and 7 males, with an average age of 19.2.

4.3.2.2. Stimuli & Procedure

The stimuli were created using the same equations as in Experiment 5 (but with the previous programming error corrected). There were 125 apartments for which participants estimated the value in the first part of the experiment. Participants were also given feedback, being told what the true rental value was, so that they could improve their estimates and calibrate to the stimuli. In the second part of the experiment participants were shown 125 different item pairs and were asked to select which they thought was more valuable.

The stimuli were presented with attributes in a horizontal row. In choice trials, one item was at the top of the screen and the other at the bottom. The position of the attributes remained the same throughout, so participants quickly learnt which attribute would be present in each location on the screen. This means that when participants fixate on a piece of information we know that they intended to look at it and are not simply exploring the information to find the current location of the information they are actually interested in. The order of stimuli was not counterbalanced between participants, potentially leaving confounds with screen position (however these are minimised by individual differences in attribute weighting).

4.3.2.3. Eye-Tracking

Gaze position was measured using an SMI RED II with monocular sampling at a rate of 50Hz. Participants were seated with a chin rest at 50 cm from the screen and 45cm from the eye-tracker sensor.

Fixations were defined as any period of at least 100ms where gaze remained within a 50 pixel radius of the mean position of the fixation. A tolerance of up to 40ms missing data during a fixation was also allowed, providing that the gaze position for present data never left the fixation radius. This resulted in an average of 76.5% of all valid gaze location readings being classified as a fixation. Areas of Interest (AOIs) were defined as rectangles 155x155 pixels centred on each number. This resulted in 90.2% of fixations falling within an AOI. These classification rates are comparable to other experiments using similar methodology.

4.3.3. Results

4.3.3.1. Behavioural results

Participants choices showed good calibration to the task environment, with the more valuable item selected on 79.5% (SD = 13.4) of trials. Participants took an average of 4.83 seconds (SD = 1.74s) from trial onset to response. Cluster corrected correlation also showed that participants responded more slowly when rental values were more similar ($r = -0.186$, $p < 0.001$). When correlations were performed separately for each participant a t-test comparing all r-values to zero shows that the effect is also robust across individuals ($t(23) = 5.66$, $p < 0.001$).

A modified regression equation was then used to estimate the relative importance participants placed upon each attribute. A standard regression was not appropriate because this would require the same attribute for the top and bottom items being entered as separate parameters. However, it is not reasonable to expect a particular attribute to have different weightings simply because it was presented on the top or bottom of the screen. The appropriate equation is very similar to that of standard regression but applies the same beta weighting to attribute values from item one and item two.

Equation 14

$$Y = B_0 + B_1 Beds_1 + B_2 Baths_1 + B_3 Sq'Ft'_1 + B_4 Crime_1 + B_5 Distance_1 - B_1 Beds_2 - B_2 Baths_2 - B_3 Sq'Ft'_2 - B_4 Crime_2 - B_5 Distance_2$$

This can then be rearranged to the following:

Equation 15

$$Y = B_0 + B_1(Beds_1 - Beds_2) + B_2(Baths_1 - Baths_2) + B_3(Sq'Ft'_1 - Sq'Ft'_2) + B_4(Crime_1 - Crime_2) + B_5(Distance_1 - Distance_2)$$

This allows for estimation of the beta weights using a standard logistic regression (Table 4.3). We wanted to test whether participants were successfully extracting the weighting functions used to calculate the true rental values given as feedback in the first part of the experiment. Therefore, the same analysis was performed but using the true rental values to determine the preferred item. This showed a very similar pattern to participants' responses. Bedrooms and bathrooms were consistently underweighted, as they were in Experiment 5. However, participants' behaviour appears reasonable and there is no reason to think this has any impact upon further analyses.

Table 4.3 Revealed weighting of each attribute accross participants demonstrates that despite relatively good fit, participants are slightly underweighting bedrooms and bathrooms.

	Bias	Beds	Baths	Sq'Ft'	Crime	Distance
True Rent	0.02	0.52	0.35	0.13	-0.22	-0.3
Estimated Weights	0.01	0.42	0.23	0.17	-0.18	-0.33
Significance of difference	$t(23) = -0.66$ $p > 0.05$	$t(23) = -2.79$ $p < 0.05$	$t(23) = -5.83$ $p < 0.001$	$t(23) = 1.54$ $p > 0.05$	$t(23) = 1.71$ $p > 0.05$	$t(23) = -1.26$ $p > 0.05$

4.3.3.2. Categorizing Individuals' Behaviour

As in Experiment 5, participants were categorized according to which model out of WADD, Dawes rule and DbS, best predicted their choices. WADD predicted choices with an overall accuracy of 86.8%. When split by trial type, 89.1% of Matched trials were correctly predicted and 85.2% of mismatched trials. As in Experiment 5, Dawes rule performed close to chance, correctly predicting 50.3% of all choices, and again had a significant split by trial type with 87.2% accuracy on matched trials but only 25.7% for mismatched trials. DbS was again worse than WADD, predicting 72.8% of choices, with 85.8% accuracy for matched trials, but only 63.9% for mismatched.

When calculated across all trials, WADD was the most accurate predictor for all 24 participants. This was also the case when the analysis was restricted to mismatched trials. For matched trials DbS was the best model for 2 participants, Dawes for 2 and 15 were again classified as WADD responders. The other participants were equally well explained by WADD and Dawes rule.

DbS was then calculated allowing attribute weightings to vary freely within the model. This increased its accuracy to 85.1% across all trials. It also reduced the effect of trial type relative to the unweighted DbS model: Accuracy was 88.5% for matched trials and 82.8% for mismatched. When calculated across all trials the model was the best fit for 5 participants, with the other 19 still best explained by WADD and this difference was significant across all participants ($t(23) = 3.98$, $p<0.001$). For mismatched trials, 3 participants were classified as DbS responders, 18 as WADD and 3 were equally well explained by both. Across all participants, WADD is still significantly more accurate ($t(23) = 4.16$, $p<0.001$). For Matched trials the results are less clear cut. WADD is still the best predictor for 9 participants and Dawes rule is for 2. Of the remaining participants, only 4 are classified as DbS responders. Three participants responses are predicted equally well by WADD and DbS, 3 equally well by Dawes and WADD, 1 by Dawes and DbS. There are also 2 participants for whom there is a three way tie. Paired t-tests revealed that for matched trials there was no significant difference between the accuracy of Dawes rule and DbS ($t(23) = 1.70$, $p>0.05$), nor between DbS and WADD ($t(23) = 1.36$, $p>0.05$). WADD is still significantly more accurate than Dawes rule ($t(23) = 2.11$, $p = 0.045$), but this effect is weak and does not survive correction for multiple comparisons.

4.3.3.3. Distribution vs Weighting

As in the previous experiment, crime had a linear effect upon value but a cubic distribution function whereas distance had the opposite linear and cubic patterns. Unlike the previous experiment, the values on either scale were completely independent. This allowed for a number of tests examining how individuals used the relative weighting and distributions in their choices. By entering quadratic and cubic terms for different attributes into the regression equation it was possible to test whether they add significant explanatory power. Thus, is there a significant quadratic or cubic component in individuals' use of stimuli values?

To assess the relative influence of each attribute and its quadratic and cubic components, a Bayesian Inference Criterion was calculated. The BIC is a poorness of fit measure and applies a penalty for each additional parameter entered into the model. Therefore a reduction in BIC value indicates that the additional parameter(s) significantly reduce the error in the model and that this is not simply due to the additional free parameter(s) fitting noise. One BIC was calculated for the difference between the full cubic model and the quadratic model and another BIC for the difference between the linear and quadratic model.

Table 4.4 shows that both crime and floor space had a moderate quadratic component, but distance showed moderate support for the linear model. Only crime showed a significant cubic effect and this effect was very large. The linear distribution of distance means that the BIC favours the linear model, despite its cubic effect upon value. This supports DbS, suggesting that a skewed distribution has a significant effect upon choices, whereas an identically skewed weighting function does not. Plotting the weighting functions estimated by the model shows that participants weighting functions are a close visual match for the cumulative frequency plots (Figure 4.4).

For choices:

Table 4.4 The difference in BIC when additional terms are added to the model. A negative BIC difference indicates that the additional parameter adds significant explanatory power to the model

	SQFT	crime	Distance
Quadratic - Linear BIC	-4.79 (10.9)	-2.05 (2.78)	1.43 (2.05)
Cubic - Quadratic BIC	17.62 (6700)	-13.5 (852)	3.89 (6.98)

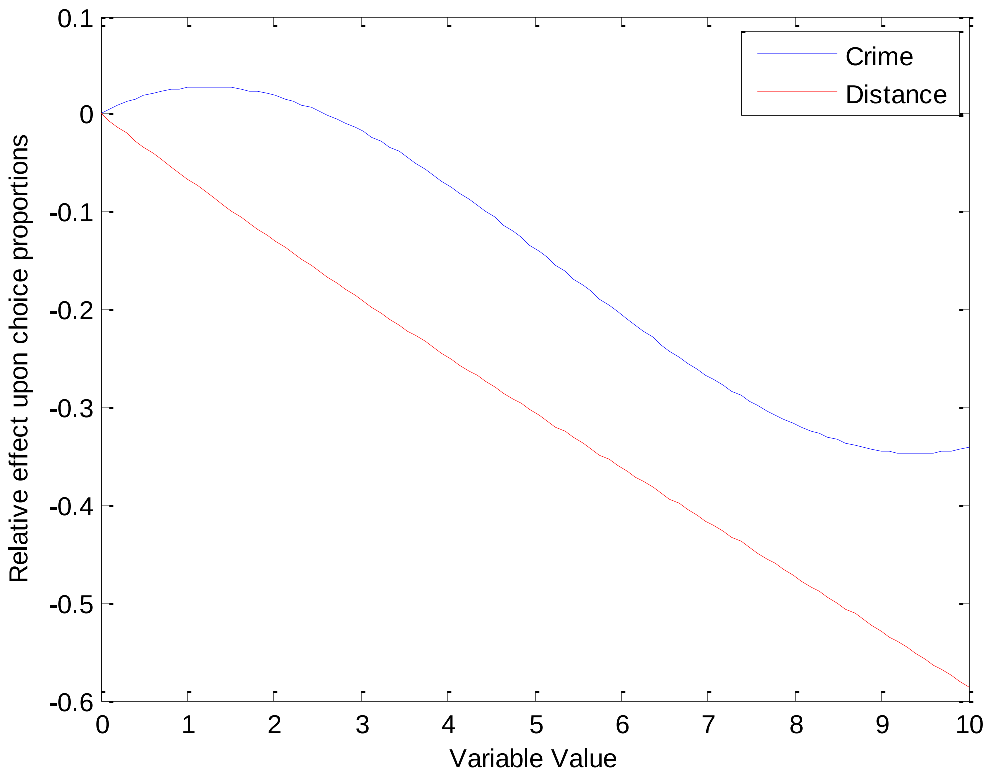

Figure 4.4 The mean recovered weighting functions which participants applied to the crime and distance attributes when making their decisions

4.3.3.4. Eye-Tracking Results

4.3.3.4.1. Gaze Cascade

Evidence accumulation models including drift diffusion often predict that individuals will attend more to the item they subsequently choose (Busemeyer & Townsend, 1993; Krajbich et al., 2010), as do models of feedback loops (Shimojo et al., 2003; Simion & Shimojo, 2007). Therefore the first question is whether participants look more often towards their preferred item. Across all trials and participants, 53.8% of fixations were towards the preferred item. A bootstrapping analysis was conducted where the chosen item on each trial was randomly assigned on each iteration. This showed that although the effect was small it was highly reliable (p<0.0001). A further prediction of drift diffusion is the gaze cascade: do participants begin to look more towards their preferred item over time? To examine this, the proportion of fixations directed towards the preferred and non-preferred items were calculated in the period leading up to the decision. Figure 4.5 reveals that the difference did indeed increase over time. There was a general trend for several seconds before response and then a steep increase immediately prior to decision. Bootstrapping analysis was again used, this time to test the significance of the difference at each time point. The chosen item was randomly assigned for each trial on each bootstrap iteration. Points where p <= 0.01 are indicated in Figure 4.5.

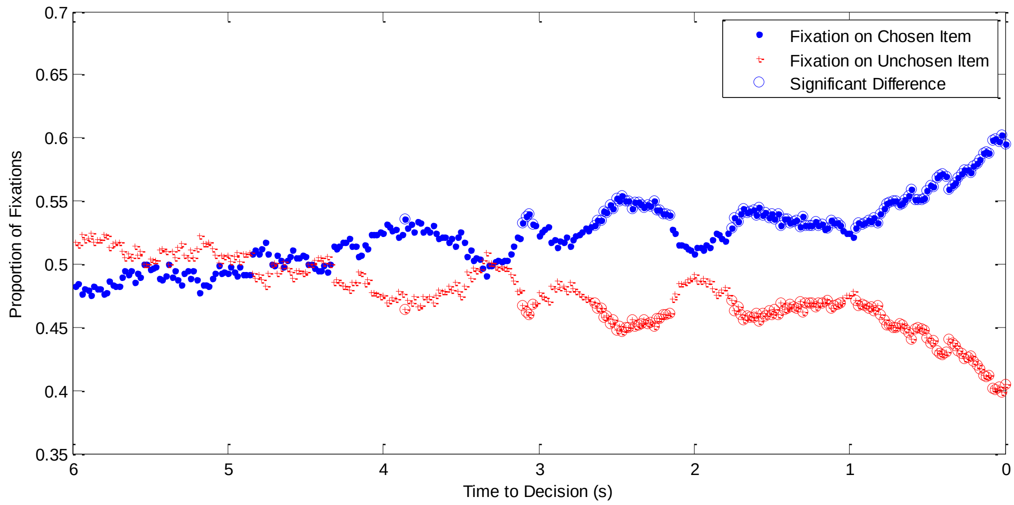

Figure 4.5 proportion of trials where fixations were directed at the preferred and non-preferred items in the lead up to response.

4.3.3.4.2. Attribute Weighting and Looking Patterns

The next question was whether participants were looking more at the information they weighted most highly when making their decisions. To examine this, the number of fixations on each attribute was used to calculated fixation proportions for each participant. These proportions were then correlated against the behavioural weightings revealed by participants' choices (as described above). This revealed no significant correlation ($r = 0.097$, $p = 0.29$). One potential reason for this is that the relative importance of each attribute varies across trials depending upon the item values. For example, on some trials each apartment has the same number of bedrooms, meaning the attribute is not helpful in differentiating between the two. It could be that participants are directing their attention towards the attribute which best differentiates the two items on that particular trial.

An additional correlation analysis was conducted, but this time the attribute values were multiplied by participants (non-standardized) elicited weightings before the difference between them was calculated. This produces a measure of how well each attribute differentiates between the items on each trial. The correlation revealed a very small yet statistically significant negative correlation ($r = -0.017$, $p<0.05$), suggesting that participants were in fact looking less frequently at the information most important to their decisions. However, when separate correlation analyses were performed for each individual, only one showed a significant correlation ($r = -0.157$, $p < 0.001$) and when this individual was removed from the overall correlation, the effect disappeared ($r = -0.009$, $p = 0.25$). This suggests that one individual was using a different strategy to the others in this experiment and that the majority of individuals did not attend more to information depending upon its importance.

Participants did not look proportionally more at influential information across the entire trial duration. However, it is possible that the effect only becomes apparent as individuals near the decision point, essentially a gaze cascade towards more influential information. Drift diffusion predicts that the drift should be steeper when an individual is attending to an attribute where there is a larger difference between the two items. This means it is most likely that a threshold will be passed and a decision made when the most influential attribute is being attended to. Therefore, at the final fixation participants should be attending most frequently to the most influential information. To test this each attributes' relative influence on each trial was first calculated as described in the previous paragraph. This allowed for each attribute's influence upon a decision to be calculated for each trial and then placed in a rank order of influence. Gaze towards the most and least influential attributes was then examined in the lead up to the decision point (Figure 4.6). Unlike with fixations to items, this shows no obvious pattern. Bootstrap analyses confirm that at the time of the response there is no significant bias to any attribute; in fact 1s prior to response it is the least influential attribute that is attended to significantly more than average.

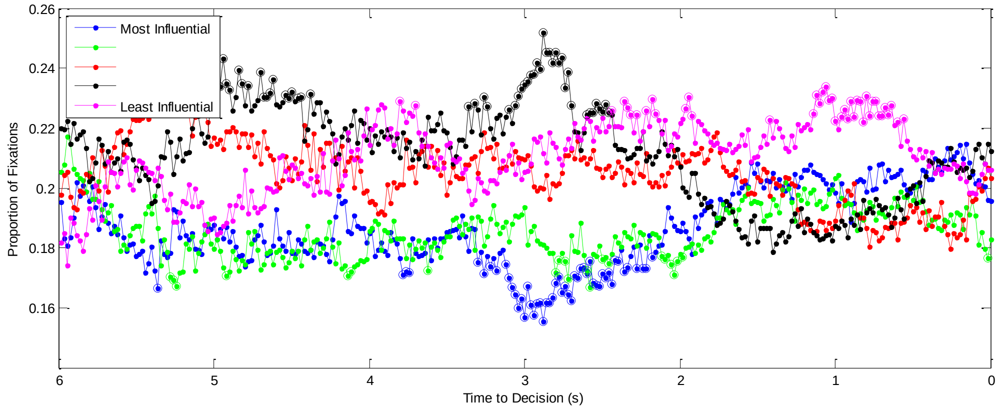

Figure 4.6 Proportion of trials where fixations were directed at the most or least influential attributes in the lead up to response.

4.3.3.4.3. Fixation Durations

A further prediction of drift diffusion models is that the final fixation before a decision will be significantly shorter than others. This is because as the accumulator crosses a decision boundary the fixation is interrupted. Contrary to some previous findings (Krajbich et al., 2010), we find participants final fixations were actually significantly longer (Figure 4.7).

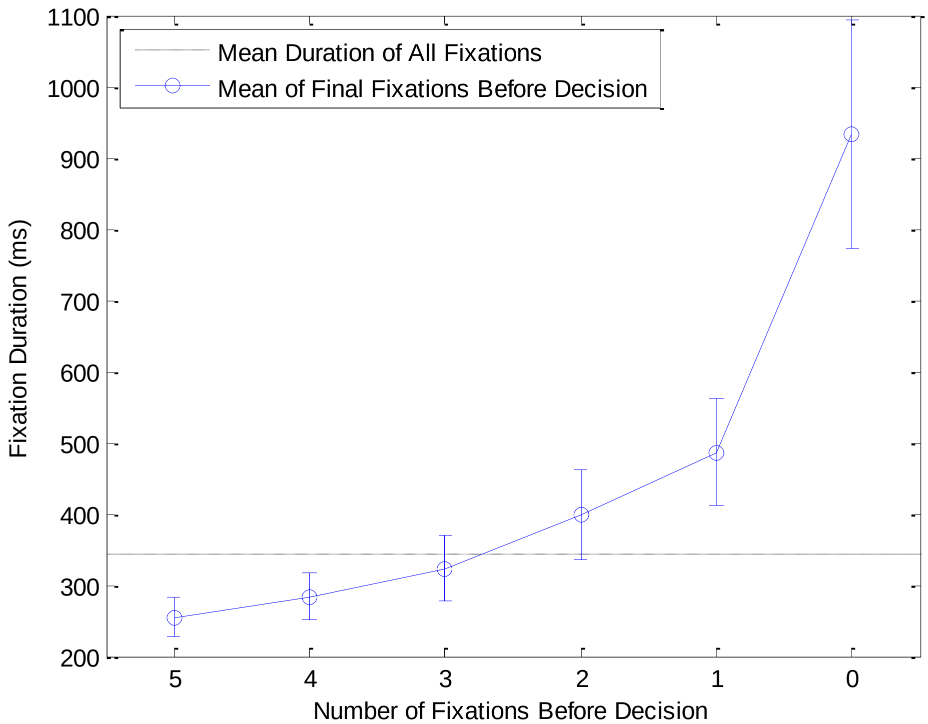

Figure 4.7 Duration of final fixations before decision.

4.3.3.4.4. Model Fitting

Other papers which report modelling of drift diffusion use iterative simulations with random error terms and an assumption of random switching of attention between items and attributes (Busemeyer & Townsend, 1993). Even experiments that have measured visual attention on each trial have then gone on to model the data by drawing random fixation durations from all those measured during the experiment. As we have continuous eye-tracking data for all participants we chose to constrain the model according to the actual attention durations on each trial, for each individual. This provides a test which is deterministic and much better constrained in accordance with the model's predictions. One consequence is that as the attention durations are determined, so are the reaction times, meaning the model can only be used to predict choices.

A logistic regression was used to efficiently calculate the final resting state of the accumulator and the best fitting parameters for each participant. For the first model the accumulation of information was assumed to be entirely independent of the unattended item. This is essentially the same as setting θ to zero.

Equation 16

$$B_A V_{A\,top} D_{A\,top} - B_A V_{A\,bottom} D_{A\,bottom}$$

Where B is the estimated weighting individuals place upon attribute A, V is the value of a particular attribute for the top or bottom item and D is the total duration of time spent attending to that information on a trial. Individuals were modelled separately to account for individual differences in the weighting of attributes. The model appeared to perform well, correctly predicting 74.1% of trials.

The next step was to incorporate the effects of attention reported in (Krajbich et al., 2010). This was achieved using a modified version of Equation 16:

Equation 17

$$\left(B_A V_{A\,top} D_{A\,top} - \theta B_A V_{A\,bottom} D_{A\,top}\right) - \left(B_A V_{A\,bottom} D_{A\,bottom} - \theta B_A V_{A\,top} D_{A\,bottom}\right)$$

Allowing θ to vary within the model revealed a mean value across participants of 0.25, which indicates a relatively strong effect of attention. However, there were significant individual differences. The standard deviation was 0.28 and the range of values covered the entire scale, including the imposed limits of 0 and 1. Overall, the model did provide a modest improvement, with 76.3% of decisions correctly predicted. To ensure that this improvement was not simply due to the additional free parameter a Bayesian Inference Criterion (BIC) was calculated for each individual model. Importantly, it also incorporates a penalty for additional free parameters and allows a direct test between both nested and non-nested models. In this case, the BIC confirmed that the improvement in fit is not simply due to free parameters (Table 4.5). The BIC provides a parsimonious method for comparing the drift diffusion models to baseline measures and conceptually different models. The most obvious is WADD as it was the best performing model in Experiment 5 and makes no account for attention. Not only did this model show much better predictive accuracy (86.8%), but also a better BIC score. This raises the question of whether relative attention has any effect upon decisions. To examine whether fixation proportions were predictive of decisions an additional model was calculated using only the number of fixations on each attribute of each item. This model correctly predicted 67.9% of choices, well above chance. However, the gaze cascade results reported above suggest that this may be due only to additional fixations upon the chosen item, with the attended attribute being irrelevant. When the model was simplified and the number of fixations upon each item used as the sole predictor, ignoring which attribute was attended, the model still performed above chance (65.0% accuracy). Furthermore the BIC revealed that the small improvement found by the full fixation model was due only to the additional free parameters.

To assess whether the predictive power of fixation proportions is independent of stimuli values, these were added to the original WADD model as additional parameters. Adding only item fixations resulted in next to no change in predictive power. Adding the full fixation model, also separating out fixations to different attributes as different predictors, slightly increased the percentage of choices predicted. However, once again the BIC revealed that this is simply due to the large number of additional free parameters.

Table 4.5 Performance measures of alternative models. BIC is a poorness of fit measure which allows for comparison of models with varying numbers of free parameters and assumptions

Model	% Accuracy	Bayesian Inference Criterion
WADD	86.8	1246
Item Fixations & WADD	86.7	1246
Attribute Fixations & WADD	89.0	1411
Attention Weighted Drift Diffusion	76.3	1889
Simple Drift Diffusion	74.1	1898
Item Fixations Only	65.0	1971
Fixations Only	67.9	2127

4.3.4. Discussion

We report the results of a multi-attribute choice experiment conducted with concurrent eye tracking. Results were fitted to drift diffusion models of choice using visual fixations as a proxy for attention. The results show that whilst a number of findings from previous single-attribute choice experiments are replicated, drift diffusion models perform significantly worse than simpler models which do not account for attention. We also find that participants direct their attention more towards their preferred item over time; however they do not direct attention towards more influential information or the information on which they base their decisions. This means that attention based accumulator models will inevitably assume improper weighting of attributes.

Therefore the best performing model tested here is a simple WADD model using only stimuli values and behavioural responses.

The experiment also calls into question the methods used to model previous drift diffusion experiments, particularly those using eye tracking. The largest difference is that reaction times were determined and the model used only choices to estimate the free parameters. Therefore the method of modelling employed here only estimates the final resting state of the accumulator at the point a decision is made. It does not explicitly model the accumulator over time and does not use a noise term or provide a distribution of responses. The reason this method is more applicable is that it allows for constraining the attention parameters of the model using the actual recordings of fixations during each specific decision. It seems a strange decision to lose the explanatory power provided by real time eye tracking in favour of an assumption of random sampling. This method is still in line with DDM, we are merely constraining free parameters with empirically recorded measures. One possibility is that visual attention is not an appropriate proxy for the attention terms within drift diffusion models. But if one accepts this then one must also accept that eye tracking as a tool has very limited potential when modelling decisions and reassess previous findings (Krajbich et al., 2010; Orquin & Mueller Loose, 2013; Philiastides & Ratcliff, 2013). It also fails to explain the reliable phenomena which are found (Shimojo et al., 2003).

An additional issue with the modelling of this particular version of DDM is the range of individual differences in estimations of the attention parameter. Large individual differences are nothing new in decision research (one need only think back to Experiment 5), but it is difficult to discern any sensible hypothesis for different people having such wildly different values in this particular instance. If attention bias is a real phenomenon at the population level then one would expect individual estimates to form a normal distribution of values around a mean. In fact what is found is a cluster at either extremes of the scale and others randomly distributed in between. This is more indicative of a mathematical model fitting noise. It is true that the BIC shows an improvement over the unconstrained model despite the additional free parameter, but the effect size is small. One must also be very careful of modelling methods when there are such large differences. In the modelling performed here individuals were modelled separately, with an attention parameter calculated for each. Previous experiments have found similar individual differences but then collapse across individuals, using the average value for further modelling.

The experiment also corrected the programming error of Experiment 5, allowing for complete modelling of participants' weighting functions and use of attribute information. This revealed that the weighting function extracted from responses matched the distribution of the scale as predicted by DbS. Crime, the attribute with a cubic distribution function but linear effect upon rental value, showed a cubic effect upon choice proportions. However, distance had a linear relationship which matched its linear distribution and this was despite its true effect upon value being cubic.

The cubic function cannot be explained in the same manner as in the previous chapters. In the single item valuation tasks of Chapter 3 the cubic function was in the opposite direction to the one found here. The extremes were over-weighted with a plateau in the mid-range of the scale. This was the opposite to that predicted by rank order and therefore by DbS. However, in this experiment the cubic shape shows a steep curve in the central section of the scale but plateaus at the under-represented extremes.

When DbS was explicitly fitted to behavioural responses, the results were the same as in Experiment 5. It performed poorly when attributes were not weighted, then when weightings were estimated, its predictive accuracy was not different to that of WADD for matched trials, but it was significantly worse for mismatched. Although suggestive, the finding comes with the same caveat: the stimuli were specifically controlled to keep the difference in WADD valuations for paired items as similar as possible for matched and mismatched trials. Even when the uncontrolled matched trials are included the ratio of WADD differences between matched and mismatched trials is only 1:1.25. However, for DbS this is 1:2.4. Therefore, DbS makes disproportionately strong predictions, all in the opposite direction to Dawes rule, simply because the stimuli were specifically controlled for differences in WADD predictions.

One of the most robust findings of previous eye tracking studies has been the gaze cascade (Shimojo et al., 2003). Immediately prior to making a response, individuals look more towards the item they subsequently choose. There are two prominent hypotheses regarding this phenomenon. Drift diffusion models suggest that the final fixation is more likely to be directed towards the chosen item because evidence should be accumulating for the attended option as the accumulator crosses the decision barrier (Krajbich et al., 2010). The second hypothesis is that there is a feedback loop from reward processing to attention (Simion & Shimojo, 2007). In this experiment preferential looking towards the chosen item begins several fixations before response and shows a relatively gradual trend over time. A drift diffusion account predicts that only the final fixation should show this pattern, whereas a feedback loop supports a gradual build up. Furthermore, the drift diffusion account predicts a similar cascade effect towards more influential attributes but we find no such effect.

It seems surprising that attention is not driven towards the more useful or influential information over time. This has significant implications for the feedback loop, suggesting that values are calculated for each item as a complete whole. The fact that attention does not tend to be directed towards more influential information over time suggests that this loop has no access to more fine grained information including that at the level of individual attributes. Findings from other disciplines suggest that this phenomena is somewhat unique to value based choices, as individuals were more likely to recall items they had fixated on for longer and to fixate longer upon more informative or task relevant objects in a complex visual scene (Henderson, Weeks Jr, & Hollingworth, 1999; Loftus & Mackworth, 1978). Therefore we do not propose that individuals fail to attend more to important information in general. Simply that the calculation of importance is calculated at the level of the item, not the attributes.

Contrary to the predictions of several models of decision making, visual attention does not drive preferences. We argue that the reverse is in fact true. The best predictor of choices is the WADD model which takes no account of visual attention. Conversely, the worst performing models are those which use only attention as predictors. Crucially, combining these models using additional parameters does not improve the performance beyond that of the original simple model . The predictive properties of visual attention patterns are collinear with that of the attribute values and are not an independent or orthogonal predictor. If visual attention alone biased decisions towards an item then it should be an independent predictor and measures of attention would provide additional explanatory power. Therefore we must conclude that biases in visual attention are driven by how an individual values the item and its attributes. Put simply, individuals look more at the option they think is better; individuals do not think an item is better simply because they have looked at it more.

As in Stewart et al (under review), we find that final fixations are significantly longer than mean duration. This is contrary to Krajbich et al and to the drift diffusion model in general: the crossing of a decision threshold should terminate a fixation early. Although there is little information at present, it seems likely that this effect is due to task complexity. Krajbich et al (2010) used a relatively simple, single attribute decision between foods. Stewart et al (under review) used financial gambles and this experiment used five attribute items. There is an increase in complexity in each just as there is a concomitant increase in the duration of final fixations between the three studies. Evidence of previous studies investigating the effect of working memory load on fixation characteristics also provides support for such a hypothesis (Gould, 1973).This would suggest that once individuals feel they have collected enough information to make a decision, they stop attending to attributes and consider their decision before responding. The gaze cascade and revisiting of information suggests that this only happens when a preference already exists, so this consideration period likely serves as a final check before response.

4.4. Chapter Discussion

The experiments in this chapter demonstrate that for the overwhelming majority of participants, choices are best described by a weighted additive model. DbS proves to be a comparatively poor predictor of individuals' behaviour, when attributes are not weighted and equal to that of WADD when weightings are applied. The results also suggest that there is no fundamental or qualitative difference in the cognitive processes employed during valuation and choices. However, they show that the interpretation of individual attribute scales and participants' weighting of information does change. Experiment 5 shows that differences between items can be accentuated or camouflaged depending upon the distribution of attribute values. These effects support the evaluability hypothesis (Hsee et al., 1999) of preference reversals and further support the notion that WADD is used for both choices and valuation. Experiment 6 goes further by identifying the shape of participants' weighting curve and their interpretation of attributes with skewed distributions. The results show that participants' use of attribute information is dictated by the shape of their distribution and the resulting rank order. The true shape of the effect of an attribute upon item value is not represented in individuals' choices.

Despite the finding of rank order encoding on individual attribute scales, DbS performed poorly when used to predict the ultimate choice. An obvious suggestion would be that this is a result of DbS's simplistic, unweighted additive integration of multiple attributes. However, the model's performance was still poor when the attribute weightings were allowed to vary freely within the model. This is surprising as the rank order characteristic of participant decision making is something which can only be captured by DbS, and not by the linear WADD model which often outperformed it. An interesting consideration is the effect of matched and mismatched trials. WADD consistently outperformed DbS on mismatched trials, but there was no significant difference between them on matched trials. This is because the stimuli were specifically created in order to control for the size of WADD's predicted preference in each of the two trial types. The same constraints were not applied with respect to DbS's predictions, meaning that DbS made much stronger predictions in the mismatched trials than the matched.

Interestingly, for an asymmetry in DbS predictions to be affecting model accuracies in the manner described above, the Dawes rule must be manipulating choices in some manner. When examining Dawes' predictive accuracy in different trials, the most striking finding is that the performance in mismatched is significantly below chance but it is close to the performance of WADD in matched trials. This would suggest that the rule gets all its predictive power from correlating either negatively or positively with WADD, but then there would be a perfect reflection in accuracy between trial types. What is actually found is that the two accuracy rates sum to over 100 in both experiments. Furthermore, the accuracy of WADD is also significantly affected by trial type, performing worse when it makes opposing predictions to Dawes rule. This means that rather than any one of the models tested providing a parsimonious explanation of participants, there are multiple factors being considered. Therefore in task environments where the effect of Dawes rule is controlled with respect to DbS' predictions rather than those of WADD, it is likely that the models' relative performance will be very different, at least on trials where they make opposing predictions to Dawes.

The results of eye tracking in Experiment 6 demonstrate drift diffusion models based upon visual attention cannot predict multi-attribute decisions as accurately as simple behavioural models. Not only are direct comparisons between these models unfavourable for DDM, but many of the attention and visual fixation effects they predict are not found or exhibit a significant effect in the opposite direction. The most fundamental and surprising of these is that individuals do not attend more to the most influential information. This has significant implications for models which rely upon effects of attention and assume that the interpretation and weighting of information is dictated in some way by visual fixations (Busemeyer & Townsend, 1993; Krajbich et al., 2010).

One possibility is that the attention switching inherent in such models is covert and this attention switches between attributes which have already been viewed and are now store d in memory. Indeed, there is good evidence of an interaction between memory load and fixations, with individuals relying more on re-attending information when memory load is high (Droll & Hayhoe, 2007; Just & Carpenter, 1976). However, this still cannot explain the results of this experiment as it still predicts a correlation between visual attention and importance of information. The only difference is that memory capacity would attenuate the effect. Even allowing for such a model of covert attention switching, the results here suggest that eye-tracking is an inappropriate methodology for assessing drift diffusion models. Either way, the results call for a reassessment of previous eye tracking experiments which model drift diffusion, particularly those which use single-attribute items.

An alternative explanation for the visual attention effects is the hypothesised feedback loop between saccade/attention planning and reward sensitive neural systems (Shimojo et al., 2003; Simion & Shimojo, 2007). As one item begins to be preferred, it is preferentially attended to, which means they attend to more evidence in that item's favour which i n turn makes it even more strongly preferred and so on. The results of Experiment 6 support this model as the gaze cascade effect is not confined to the final fixation. Instead it shows a more gradual development over time. The particularly interesting finding here is that the feedback loop is blind to attribute level information. There is no cascade towards the more informative attributes or information on each trial. The gaze cascade itself is only found for complete items. Therefore, the inputs to the fe edback loop are necessarily the current value estimates of the options. This in turn suggests that the neural systems processing reward represent only store a persistent estimate of the items' overall value, not the values of individual attributes. Chapter 5 examines the neural correlates of reward processing in more detail.

5. Chapter 5

5.1. Chapter Introduction

Just as recent years have seen an upsurge in attention to comparative models such DbS in the behavioural literature, the same is beginning to happen in neuroscience and neuroeconomics. Research is instead moving towards finding a truly explanatory model that describes the underlying process or neural systems that *cause* patterns of responding (Louie & Glimcher, 2012; Vlaev, Chater, Stewart, & Brown, 2011; Weber & Johnson, 2009). Much of this work in neuroeconomics has focused upon the importance of previous experience and stresses the relative evolutionary importance of action choice over the comparatively very recent requirement for calculating an isolated scalar valuation, i.e. a financial judgement (for a review see Seymour & McClure, 2008; or Vlaev et al., 2011).

In this chapter, two fMRI experiments examined neural responding to financial rewards. These experiments tested the predictions of rank order encoding and examined the qualities of the task environment which dictate the sample of previous experiences recalled. The results of these experiments revealed novel findings relating to neural encoding of value. In Experiment 9 these neuroscience findings were then tested in an analogous behavioural task which successfully demonstrated a novel manipulation of utility curves as well as providing a cross-modality replication.

5.2. Experiment 7

5.2.1. Introduction

The predictions of DbS map very intuitively onto neuroeconomics and neural systems. It is a widely accepted finding that there are value responsive regions which show greater activity to higher value rewards (Knutson & Bossaerts, 2007; Kringelbach & Rolls, 2004). However, neural firing has biologically defined maxima whereas financial values can increase to infinity. This means that if there were a one-to-one ratio between neural firing and value, individuals would be unable to differentiate the relative benefits of 10pence vs 30pence as well as £1million vs £3million. Applying behavioural models such as DbS would predict that a scale is created anew for each valuation by recalling from memory a sample of items similar to the current one. The neural activity would then represent the item's rank within the recalled sample. This closely corresponds to an often held assumption that neural responding is context specific, with activity representing the difference between the current context/environment average and the current item's value (Knutson & Wimmer, 2007; Tobler, Fiorillo, & Schultz, 2005).

One source of evidence for context dependency comes from paradigms investigating what is often referred to as "menu context": so called because the tasks are analogous to choosing your preferred dish from a restaurant menu. Two or more stimuli with different pre-trained values are shown to participants and then one is selected randomly and indicated as the reward to be received from that trial. Both primate single cell recording and human fMRI have shown that activity in the Orbito-Frontal Cortex/ventral medial Pre-Frontal Cortex (OFC/vmPFC) reflects the stimuli's value relative to the other possible rewards shown at the beginning of the trial (Elliott, Agnew, & Deakin, 2008; Tremblay & Schultz, 1999). The same stimuli can elicit maximal responding when paired with less preferred stimuli and minimal responding when paired with more preferred stimuli, even though its objective value remains the same.

The vmPFC is also implicated when subjects make an active choice and select the most valuable option (Knutson et al., 2008; Rangel & Hare, 2010) and has also been shown to have a more general role in response selection and action planning (Rogers et al., 2004; Schoenbaum, Setlow, & Ramus, 2003). However, the region also responds to the reward value of a single item, when there is no choice required and no immediate alternative with which to compare (Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001). This is generally interpreted as the region being implicated in two entirely different processes: a parametric calculation of independent value and comparison of available choice alternatives (Hunt et al., 2012).

However, the predictions of DbS provide another, more parsimonious and cognitively efficient explanation: that the response elicited by a single value is also calculated by the same comparative process, but using alternatives retrieved from memory.

Many studies have used contextual manipulations that do not involve any menu context and present only one stimulus or value at a time. Some have used visual categorization cues such as shape or colour. For example, using green cards to signify gain trials and red cards to signify loss trials before they are turned over and the amount won/lost is revealed (Cooper, Hollon, Wimmer, & Knutson, 2009; Nieuwenhuis et al., 2005). In the critical comparison of $0 win vs $0 loss, greater responding to the $0 loss was found most strongly in the ventral striatum and the best possible outcomes in both contexts activated the region to a comparable degree. As a result of these studies, it has been suggested that the brain calculates an average environment value which is then used as a baseline comparison for current items (Knutson & Wimmer, 2007). This is based upon the compelling evidence that the activity in the Ventral Striatum relating to value is context dependent, but troublingly for this model the predicted baseline signal has not yet been identified. However, if the response to each stimuli is calculated by comparison with other similar or recently experienced items then this makes the response inherently context dependent, but without the need for an explicit baseline to be calculated. In fact if this baseline were found it would immediately raise the original problem of how a finite activity range represents an infinite value range; it would merely move the issue from the VS to wherever the baseline was calculated.

A question raised by models such as DbS is whether such a system of comparison would be a simple rank order, better/worse comparison or a more complex parametric comparison which is able to represent the scalar difference between alternatives. Although there is existing evidence that the vmPFC does represent absolute difference when attending to one of a pair of items (Basten, Biele, Heekeren, & Fiebach, 2010; Lim, O'Doherty, & Rangel, 2011; Philiastides, Biele, & Heekeren, 2010), it is additionally possible that this is simply a representation of the rank difference within the context of all experiment trials. Indeed, if it is a representation of absolute difference then once again we return to the question of how the infinite range of potential differences can be represented.

The experiment described here presented pictures of cash, one stimulus at a time in a manner which did not require or overtly encourage comparison between them. These were split into blocks of high and low value trials such that it was possible to examine whether value dependent regions responded in a context dependent manner solely due to recency of stimuli exposure. Furthermore, the experiment used a distribution of values which is non-linear so that it is possible to examine the pattern of responding and test whether it represents rank order or absolute financial value. The results revealed that the VS and Thalamus are strictly context dependent within block, showing similar activation to the lowest and highest value in each block. Furthermore, activity in the vmPFC and the Anterior Cingulate Cortex are not constrained by the context of block, but show a strongly linear increase across all stimuli which can be interpreted as encoding rank order.

5.2.2. Method

5.2.2.1. Participants

Research was conducted with the ethical approval of The University of Nottingham's Medical Ethics board and informed written consent was obtained from all participants. Fourteen individuals from the Nottingham area participated in the study: 9 female and 5 male, aged between 20 and 27. Participants were told they would be paid a minimum of £10 with an additional amount dependent upon their performance within the task. The lowest amount earned was £15.00 and the highest was £23.80 with a mean of £20.67.

5.2.2.2. Procedure

A variant of the Monetary Incentive Delay (MID) task (Knutson, Fong, Adams, Varner, & Hommer, 2001) was used (Figure 5.1). On every trial participants were shown a photograph of an amount of cash. After a randomized delay they responded with a rapid button press in order to win the amount previously signified. Feedback was then given, informing participants of whether they had responded before the deadline, thus winning on that trial as well as the amount that they had won or failed to win. The advantage of the MID task is that it required participants to engage in the task and gave them a vested interest in the value of the stimuli being presented, but did not confound the value with any choice or response selection, as there was only ever one response to make. Therefore any variation in BOLD signal at Conditioned Stimulus (CS) presentation is due to differences in value representation which is independent of choice and as is shown later, independent of motoric action planning.

The trials were split into two block types: a low value block – 10p, 20p and 30p - and a high value block – £5, £7 and £10. Each block contained 60 trials, with 4 blocks being presented during the experiment and blocks 2 and 3 separated by anatomical image collection. Thus there were 240 trials in total. The order of block presentation was counterbalanced across participants with half seeing the high value blocks first and the other half seeing the low value block first. The length of the response window was controlled by a 1 up 2 down adaptive staircase, resulting in an accuracy rate of ~66% for each participant regardless of individual differences in average reaction times. Participant payment was calculated by taking the outcome of a random subset of trials which would add up to £30 then dividing the total won from these trials by two. As this was added to the guaranteed £10 received simply for taking part, the average payout was designed to be £20.

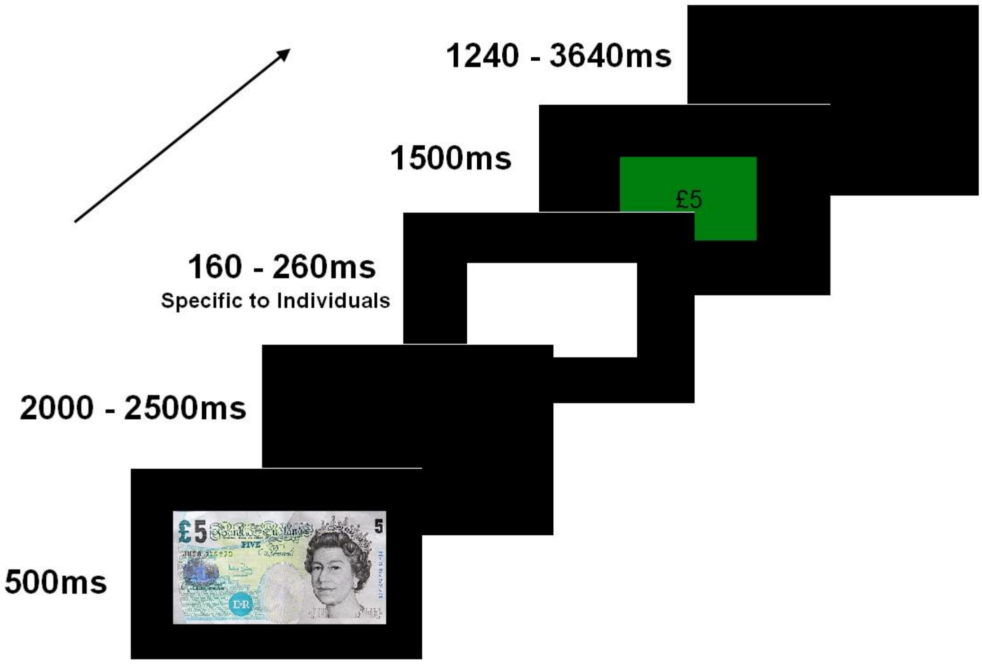

Figure 5.1 - The Monetary Incetive Delay (MID) task employed in the experiment. Note that the duration of the interval between the picture of money and the response is randomly varied to prevent anticipation and the response window itself is controlled by an adaptive staircase. The Inter Trial Interval varies to accommodate these fluctuations, maintaining the same total duration on each trial.

5.2.2.3. Scanning Parameters

Scanning was performed in a 3T Phillips Achieva scanner with 32 channel phased array head coil. To compensate for signal dropout in frontal regions a double echo, echo planar imaging sequence was employed during functional image acquisition. Previous research has demonstrated that a weighted combination of fMRI timeseries from different echo times helps combat signal dropout due to variation in peak T2* signal (Poser, Versluis, Hoogduin, & Norris, 2006).Each functional scanner run lasted 7 minutes and collected a total of 175 volumes of 36 slices for each TE using a voxel size of 3x3x3mm, TR = 2.5s, TE_1 = 20ms, TE_2 = 45ms, flip angle = 80°.

5.2.2.4. Functional Data Analysis

To ensure the accurate combination of data from both echoes realignment parameters were calculated based on images collected with the first echo sequence and were then applied identically to image sequences for both echoes. Once realigned and corrected for head motion, a weighted summation was calculated combining both echoes into one time series upon which all subsequent processing and analysis was performed. This was performed using code developed by the Sir Peter Mansfield Magnetic Resonance Centre (Gowland & Bowtell, 2007). Each voxel is weighted according to its point on the BOLD sensitivity curve, with TE's closer to the peak of the curve being given greater weight when images are combined (Posse et al., 1999). Weighted images were then transformed to MNI space using participants' anatomical scans, before being smoothed with a 5mm FWHM Gaussian kernel.

The onsets of each CS (10p, 20p, 30p, £5, £7, £10) within each block (early and late) formed the 12 regressors of interest which were entered into the first level GLM using SPM8 software to control for unmodelled error between scanner runs. A total of 24 nuisance regressors were also calculated in the same manner to encompass all values of win and lose feedback events, as well as four for button presses (one per scanner run).

Pairwise comparisons were calculated within subjects before being entered into second level random effects analyses. Comparisons were conducted for highest absolute value vs lowest absolute value (£10 vs 10p) and within block value (£10+30p) vs. (£5+10p). In addition, the highest low block value was compared against the lowest high block value (30p vs £5) as if any responding is solely context dependent then one would expect greater responding to 30p despite the large difference in absolute value. Due to the greater number of data-points in the within block contrast there was a significant difference between the power of this contrast and the other two. Therefore, to allow us to better elucidate the patterns of responding a threshold of $p<0.005$ and $k>77$ was used for the absolute value and overlap comparisons while $p<0.001$ and $k>39$ was used for within block value. These cluster thresholds were calculated using AlphaSim (Cox, 1996) such that the corrected α = 0.01. Once regions responding to value had been defined marsbar (Brett, Anton, Valbregue and Poline, 2002) was used to extract the beta weights across all conditions in order to examine the specific pattern of responding. Note that when identifying responding as representing absolute value or rank order responding all subsequent ROI and beta weight analyses are performed upon all data points while whole brain contrasts used only a subset of these conditions and trials. Furthermore, these additional analyses are conducted to test a hypothesis independent of that tested in the first level model. Both absolute value encoding and rank order encoding predict the same difference between the lowest and highest values, their predictions only differ with regards the four data points in between, i.e. the independent data points used only in the ROI analyses (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). Thus the first level model identifies all regions where the largest reward elicits greater responding than the lowest; these additional analyses serve to categorize the pattern of responding in the mid-range of this scale. The additional analyses are also performed upon ROIs identified by the within block value contrast, which uses 4 of the 6 trial types. Although in this case a larger sub-set of the data is used in the whole brain contrast, the subsequent analysis of beta weights is intended to ensure regions were responding with a similar magnitude in both high and low value blocks, thus ensuring that the hypothesis being tested is independent of that used to select the ROI. This is included as a check and to ensure completeness of information for the reader.

Correlations were performed upon extracted beta weights from each ROI for each potential pattern of responding: within block context dependency(1/3, 2/3 3/3, 1/3, 2/3, 3/3), ordinal rank(1/6, 2/6, 3/6, 4/6, 5/6, 6/6), and absolute value (0.01, 0.02, 0.03, 0.5, 0.7, 1). The within block predictions were entered only to control for the unlikely event that a context dependent area had been mis-categorized by the first-level analysis. A conservative bonferroni corrected α was used to correct for the total number of correlations performed. Where regions showed a significant correlation with two different potential responding patterns then they were entered into separate GLM's and the deviance of the models extracted. The difference between these deviances was then used to calculate a chi square statistic to test for a significant difference between them (Cohen, 2003).

5.2.3. Results

5.2.3.1. Behavioural Results

Average response times ranged from 168ms for the quickest participant to 258ms for the slowest and the overall average was 202ms (S.D. 21ms). Accuracy ranged from 46% to 67% with a mean of 59% (S.D. 5.5%). A six level one-way ANOVA showed no effect of trial value on reaction times ($F(1,5) = 1.85$). This demonstrates that differences in BOLD response are due to value calculation and are not simply a result of motoric action planning.

5.2.3.2. fMRI Results – Early Blocks

5.2.3.2.1. High vs Low: £10 vs 10p

First analysed were the results from the first two blocks where participants were not expecting the change in value range. Initially the data was collapsed across both presentation orders and responding to absolute value was examined: £10>10p. If the neural response of any area is specifically tuned to absolute financial values without relying on context or re-scaling then it should be evident here, but the only activations found were in the cerebellum and visual cortex (table 1). There were no significant clusters in the reverse analysis.

Table 5.1 Significant Clusters for Early Blocks; Highest >Lowest Value Comparison: £10>10p

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Calcarine Fissure/Lingual Gyrus/Cuneus (L/R)	8	-90	12	3096	7.889
	2	-88	-4		7.418
	10	-66	6		6.023
Cerebellum (L/R)	28	-76	-18	290	6.233
	22	-82	-18		5.912
	16	-74	-18		5.619
Occipital Temporal Gyrus (L)	-32	-72	8	280	5.642
	-34	-78	-4		4.133
	-38	-68	-16		4.032

5.2.3.2.2. Context Dependent Responding: (£10+30p) vs (£5+10p) and 30p vs £5

Contrasts were then examined which tested for context dependence that would be indicative of within-block scaling. Firstly, within block value revealed significant activation most notably within the caudate, posterior cingulate and pre central gyrus (Table 5.2). The context dependent nature of these activations is further evidenced by significant activation in the 30p>£5 (Table 5.3).

Table 5.2 Significant Clusters for Early Blocks; Within Block Comparison: (£10+30p) > (£5+10p)

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Visual Regions	-6	-90	2	5502	9.081
	0	-70	10		8.999
	26	-96	18		8.888
Posterior Cingulate/white matter (R)	20	-22	30	143	6.051
	24	-30	30		5.465
	18	-34	36		5.195
Caudate Head (L/R)	-4	6	0	122	5.590
	-10	-2	-4		5.012
Precentral Gyrus (L)	-16	-28	50	56	4.784
	-28	-22	52		4.742
Superior Parietal Lobule (R)	22	-60	48	47	4.595
	18	-54	52		4.415
Occipital Cortex (L)	-22	-60	38	45	4.520

Table 5.3 Significant Clusters for Early Blocks; Highest Low Block Value > Lowest High Block Value: 30p > £5

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Superior Parietal Lobule/Posterior Cingulate(L/R)	-4	-46	54	828	7.556
	-6	-30	34		5.182
	-10	-66	48		5.021
Precentral/Middle frontal gyrus(R)	22	-4	46	836	7.377
	26	-24	50		7.167
	38	-6	50		7.045
Middle Frontal Gyrus(R)/Middle Cingulate(L/R)	20	20	46	940	6.785
	-2	8	32		5.409
	8	18	46		5.367
Precentral Gyrus(L)	-32	-8	46	101	6.264
ACC(L/R)	-6	28	12	156	6.093
	6	34	24		5.875
	6	24	12		3.935
Middle Temporal Gryus/Occipital Gyrus(R)	54	-76	6	112	6.002
Middle Frontal Gyrus(L)	-24	30	32	99	4.317
	-28	38	30		3.594
	-26	20	30		3.162
Superior Frontal Gyrus(R)	18	44	44	77	4.281
	8	46	42		3.637
	26	40	42		3.585
Middle Frontal Gyrus(L)	-28	54	12	78	4.250
	-40	54	14		3.531

5.2.3.3. Late Block Results

5.2.3.3.1. High vs Low: £10 vs 10p

By the start of the third block participants had experienced the whole range of values. The non-linear distribution of our stimuli allowed examination of whether this responding was predicated upon a ratio scale or an ordinal scale of alternative preference as predicted by behavioural models such as Decision by Sampling. To identify the regions that respond to high values over low, regions which were significantly active in the £10>10p contrast (Table 5.4) were examined. The regions showing the most reliable activation were the ACC and vmPFC (Figure 5.2). No significant clusters were in the reverse analysis.

Table 5.4 Significant Clusters for Late Blocks; Highest >Lowest Value Comparison: £10>10p

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
ventral medial PFC (R)	6	58	-4	98	7.277
Anterior Cingulate (L)	-12	40	14	99	5.965
Inferior Frontal Gyrus (L)	-28	36	12	94	5.395
	-34	32	2		3.287
Supramarginal Gyrus (R)	30	-40	28	203	4.928
	30	-36	14		4.332
	40	-46	30		4.076
Cerebellum (R)	28	-68	-18	161	4.773
	32	-58	-18		3.948
	24	-80	-18		3.779
ventral medial PFC /ACC (L)	-10	60	8	98	4.593
	-14	52	-4		4.156
	-18	50	4		3.705
Cerebellum (L/R)	6	-68	-8	159	4.540
	-8	-84	-16		3.840
	-18	-78	-18		3.643
Supramarginal Gyrus (L)	-36	-36	16	110	4.485
	-44	-38	26		3.560
	-36	-46	22		3.444
Calcarine Fissure (L/R)	-4	-92	-4	161	4.283
	-2	-84	8		3.585
	4	-82	14		3.425
Occipital Gyrus (R)	28	-68	36	140	3.950
	34	-68	22		3.753
	30	-62	28		3.579

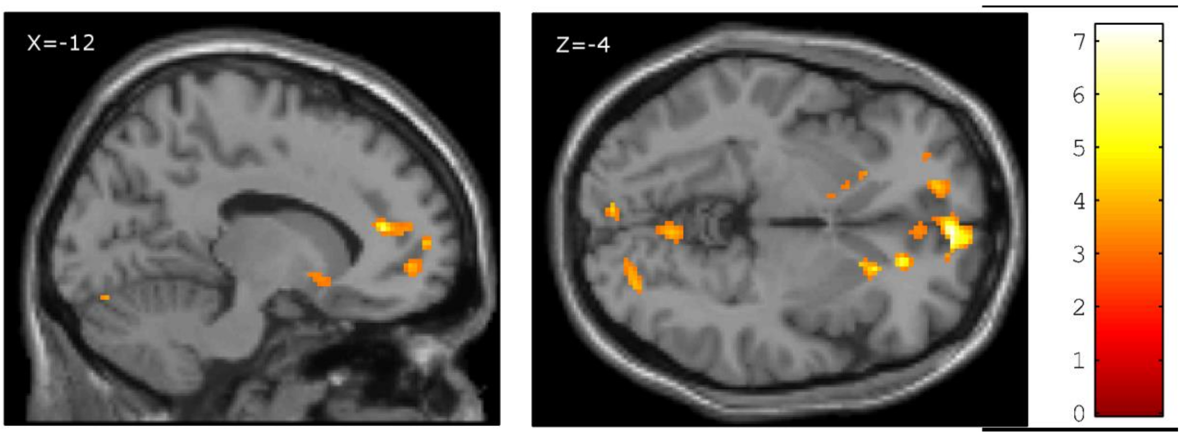

Figure 5.2- The Anterior Cingulate Cortex and ventral medial Pre-Frontal Cortex activations in the high vs. low value contrast

5.2.3.3.2. Context Dependent Responding: (£10+30p) vs (£5+10p) and £5 vs 30p

When within block responding was examined strong activations were found in the thalamus and caudate nucleus, as well as the same ACC and vmPFC regions which were active in the high vs low contrast (Figure 5.3 & Table 5.5). As in early block trials, this suggests that the thalamus and VS are responding in a context dependent manner that is sensitive to block. This is continued by the 30p>£5 contrast (Table 5.3) which found significant contextual differences in the thalamic activations (Figure 5.5). This is despite the fact that the stimulus that elicited lower responding actually had a higher objective value. There were also no effects in the reverse analysis (i.e. £5>30p).

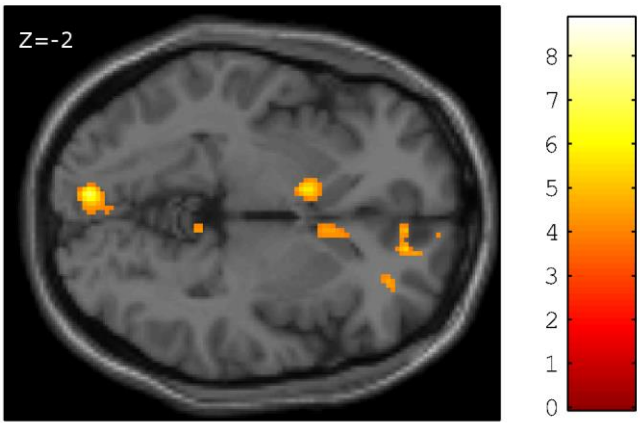

Figure 5.3 - Regions of bilateral activation within the caudate nucleus and within the ventral medial Pre-Frontal Cortex

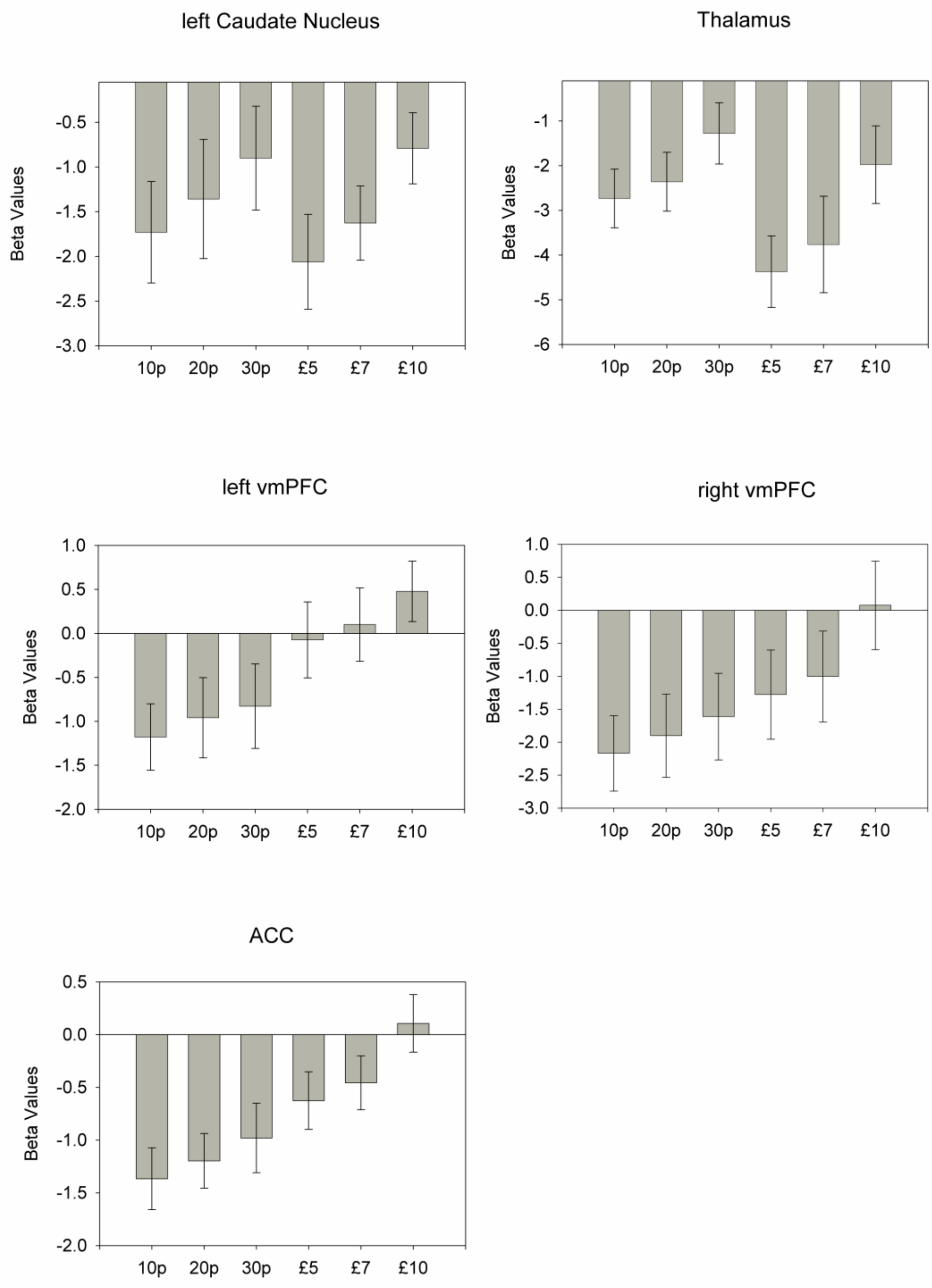

Figure 5.4 – Extracted beta weights for the latter blocks. The caudate and thalamus show local context dependency, re-scaling depending upon the values within the current block. The ventral medial Pre-Frontal Cortex and Anterior Cingulate Cortex demonstrate responding which represents a simulus' rank order within all values experienced during the experiment, independent of their true financial values.

Table 5.5 Significant Clusters for Late Blocks; Within Block Contrast: (£10+30p) > (£5+10p)

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Cuneus/Occipital Gyrus/Lingual Gyrus (L/R)	14	-96	8	1761	8.833
	-8	-90	-2		7.128
	14	-88	32		6.265
Thalamus (L/R)	4	-12	10	232	7.939
	16	6	16		5.119
	2	-24	8		4.920
Occipital Gyrus (L)	-28	-86	20	312	7.613
	-26	-78	20		6.348
	-34	-86	14		5.297
Superior Parietal Lobule (R)	32	-58	54	465	7.558
	26	-66	58		6.636
	38	-52	56		5.357
Middle Frontal Gyrus (R)	22	42	46	140	6.495
	30	40	42		5.871
	14	46	44		5.195
ventral medial PFC/ACC (R)	10	60	-6	78	6.417
	16	50	-2		5.214
Superior Parietal Lobule (L)	-36	-50	54	141	6.087
	-30	-56	50		5.939
	-42	-34	56		5.015
Lingual Gyrus (R)	8	-38	2	85	6.016
	16	-36	8		4.902
Caudate Head (L)	-10	6	-2	57	5.712
Medial Frontal Gyrus (R)	16	26	36	72	5.707
	10	34	38		4.951
	16	18	44		4.291
ACC (L/R)	8	50	32	146	5.076
	-4	42	18		4.788
	-10	48	16		4.511
Occipital Gyrus	22	-72	-14	60	5.060
Caudate Tail (L)	-18	-2	24	62	5.055
	-20	-10	18		4.297
Middle Frontal Gyrus (R)	36	0	52	58	4.887
	28	-2	54		4.320
Caudate Head (R)	8	14	-4	47	4.515
	12	24	0		4.190

Table 5.6 Significant Clusters for Late Blocks; Highest Low Block Value > Lowest High Block Value: 30p > £5

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Ventral Lateral/Ventral Anterior Nucleus (R)	8	-8	10	128	6.715
	0	-4	10		3.606
Inferior Frontal Gyrus (R)	44	46	-10	155	6.035
	38	56	6		3.628
	34	52	-10		3.297

Further analyses were performed upon all regions found to be responding to value in either the high vs low contrast or the within block contrast. A correlation was performed upon each ROI for each potential pattern of responding: within block context dependency, ordinal rank and absolute value. As stated in the method section, this analysis avoids the pitfalls of double dipping as it uses all of the data points whereas the previously applied contrasts used only a subse t. In addition, a conservative Bonferroni correction was used to control for multiple comparisons. As one would expect, significant context dependency was confirmed in the areas identified by the first level context dependency contrast: caudate ($r(84) = 0.29$, $p = 0.006$), thalamus ($r(84) = 0.28$, $p = 0.009$). Of greater interest were the regions identified in the £10>10p contrast: significant linear (rank order) responding was found in the ACC ($r(84) = 0.41$, $p<0.001$), vmPFC ($r(84) = 0.48$, $p<0.001$) and IFG ($r(84) = 0.53$, $p<0.001$). However, the results are similar when testing for value dependent responding: ACC ($r(84) = 0.4$, $p<0.001$), vmPFC ($r(84) = 0.47$, $p<0.001$) and IFG ($r(84) = 0.54$, $p<0.001$), therefore the data was entered into separate GLM models and the difference in deviance produced by them was used to calculate a chi square statistic. This reveals that all regions are actually better described by a pattern of rank order responding: ACC, $\chi(1, N = 84) = 66.4$, $p <0.001$, vmPFC $\chi(1, N = 84) = 60.6$, $p <0.001$, IFG $\chi(1, N = 84) = 48.8$, $p<0.001$. Given previous findings, it is surprising that no region is found to be representing absolute value. However, the ACC shows a strong linear relationship with rank order, while the OFC shows the same pattern in the right hemisphere and only a small deviation in the left (Figure 5.4). This is crucial for assumptions of basic human value calculation as these areas are generally considered to deal with higher level processing or the amalgamation of various low level processes (Rangel & Hare, 2010).

5.2.4. Discussion

This experiment tested some of the most basic properties of how value is represented in the brain. What role context and recent experience plays in the reward encoding of different regions and whether these regions are encoding objective absolute value, or a simpler representation of rank order within recently experienced values. It was found that the VS and thalamus are highly context dependent, with activity representing a stimulus' value relative to others in that particular block. It was also shown that the ACC and vmPFC are not affected by block, but also are not representing absolute financial value. Instead the activity in these regions represents the current stimulus' rank order within all values experienced during the experiment. No region was found with activity representing absolute financial value.

The lack of response during the early blocks in regions more generally linked with value responding may seem surprising. It can in part be attributed to participants still learning the task itself, and although this does seem to reduce power across all comparisons there are still regions such as the posterior cingulate and caudate which are already demonstrating within block value responding. Therefore, when considered alongside the responding patterns found in later blocks it seems likely that regions such as the ACC and vmPFC are not yet responding linearly because the individual is still learning the entire range of stimuli values. As these regions integrate values from a longer time period, it is also likely that values experienced just prior to the experiment are still biasing responses. Conversely, those regions which demonstrate within block responding will by definition only incorporate the more recently experienced values, hence their presence in early block analyses.

The cerebellar and visual activations found in the early blocks are not surprising. Other studies have explicitly investigated patterns of response to reward value in these regions, regularly finding increased activation to higher rewards even in very low level visual processing (Serences, 2008). This is generally interpreted as feedback from value sensitive regions with the purpose of directing attention towards more preferable and more valuable stimuli (Shimojo et al., 2003). What is confusing is that the strongest anatomical links to the visual cortex come from the caudate nucleus and thalamus (Leh, Ptito, Chakravarty, & Strafella, 2007; Platt & Glimcher, 1999). These are the regions which demonstrate the strongest context dependency in their responding to value, but the visual cortex displays the reverse pattern with significant activation only in the highest vs lowest contrast. This suggests that there is additional mediation that is yet to be described, or that there is heterogeneity in VS neuron responding which is undetectable with fMRI's spatial resolution.

The later blocks showed responding in regions more traditionally linked with value computation: ventral striatum, vmPFC, ACC and thalamus. What is interesting is the manner in which context and the value of recently viewed stimuli modulated the response patterns of these different regions. The VS and thalamus were strongly context dependent, responding according to the relative rank value of a stimulus within a block. By these later blocks participants had experienced the entire range of stimuli and although the vmPFC and ACC showed no scaling to block, they instead demonstrated a linear pattern of responding across all stimuli in the experiment. Importantly, this ordinal responding was not modified by the very non-linear distribution of actual values, demonstrating that these regions only encoded rank order preference. To our knowledge, this has not been demonstrated before and provides further support for theories which are predicated entirely upon valuation by comparison with items in memory (Stewart et al., 2006).

The finding that (unlike the lower function dopaminergic regions) the vmPFC and ACC are able to integrate stimuli/values which are contextually differentiated, supports suggestions that these areas are further along the processing chain; that they act as input integrators (Basten et al., 2010; Philiastides et al., 2010; Rangel & Hare, 2010) and response-action selectors (Hadland, Rushworth, Gaffan, & Passingham, 2003; Rushworth, Noonan, Boorman, Walton, & Behrens, 2011) respectively.

This points to a distributed, hierarchical system where lower regions are modulated by local context and those further along the processing chain integrate a wider range of information. However, these results do not suggest context independence in these regions, just that the context is far less local and integrates more information. The fact that their activity represents rank order means that they are inherently context dependent and inherently re-scale to the range of stimuli (Kobayashi, Pinto de Carvalho, & Schultz, 2010). But, the contextual boundary is defined by the experiment as a whole, rather than by task blocks. This wider contextual definition is likely the reason that fewer experiments have reported contextual effects in these regions.

It has been suggested that the ventral striatum calculates the difference between cue value and environment value: (CV – EV), (Knutson & Wimmer, 2007). However, previous experiments have not varied the environment value and the analyses have assumed a constant EV. Therefore it was not possible to dissociate whether activity is responding according to pure value or using environment value as a baseline. To our knowledge, this is the first human study which explicitly manipulates environment value during the experiment without confounding gains and losses. Although the results did show a pattern of responding in the VS corresponding to (CV – EV), it is telling that neither this nor any other experiment has found a region with activity corresponding to EV. Even if a region was found to have this pattern of responding, one would still be left with the problem that no system can represent an infinite range of values without some scaling mechanism. All that has been done is to move the problem from cue-dependent responding to environment or baseline dependent responding. If however one assumes the more parsimonious hypothesis, that there is no explicit baseline and that an individual valuation is predicated upon comparisons with similar items, then these issues are easily overcome.

5.3. Experiment 8 – Recency vs Categorization

5.3.1. Introduction

The results of Experiment 7 highlighted the effect of context in determining the neural response to a reward. However, the experiment cannot elucidate what defines those contextual boundaries. The following experiment answers at least one of these questions and in the process shows why some modelling techniques used in previous behavioural experiments were unsuccessful.

In Experiment 7 there were several potential characteristics of the task which may have caused the contextual effects found in the thalamus and VS. The first is simple recency. If the memory sampling mechanism in DbS matches the properties found in memory research then one would expect that the most recent experiences are more frequently sampled. Therefore the simple fact that the trials were separated into high and low blocks causes the recent within block experiences to be preferentially sampled. Alternatively, it could be a simple categorization effect. The use of cash photographs means that there was an inherent visual difference between the coins of low value blocks and the notes of high value blocks. Furthermore, the large difference between the pence and pound values offered in the different blocks may have indicated a qualitative change in context.

Several previous studies have found an effect of context between blocks, but these have also employed some form of salient categorisation cue. These include gains versus losses, squares vs circles, different stimuli colours and certain vs probabilistic payouts (Cooper et al., 2009; Knutson et al., 2008; Nieuwenhuis et al., 2005). There has never been a pure test of the effect of recency. That is, do the phenomena of memory research such as the Ebbinghaus curve generalise, does recency alone cause the contextual valuation effects?

This experiment uses the same task as Experiment 7. However, different stimuli are used in order to eliminate the effects of psychological context: The values used in the low value block remain the same but those in the high block are reduced to 30p, 40p and 50p. This removes the effect of value difference causing categorization. It also introduces an overlap value, as 30p is present in both blocks. This allows for a simple analysis when testing for context dependency as one can merely compare responding to the same value in the two different contexts. In addition, photographs are no longer used. Instead geometric shapes are used in order to eliminate visual contextual cues. Each shape has a unique associated value and participants are pre-trained on the task outside of the scanner to ensure they learn these values beforehand.

5.3.2. Methods

5.3.2.1. Participants

Eleven individuals (4 males and 7 females) from the Nottingham area participated in the study, aged between 19 and 34 (with a mean of 25). One additional participant completed the study but their data had to be discarded due to data corruption. Each received a £3 inconvenience allowance for the behavioural training session and an amount dependent upon their performance in the scanning session that was weighted to average £10.

5.3.2.2. Monetary Incentive Delay Task

The same Monetary Incentive Delay (MID) task (Knutson et al., 2001) was used as in Experiment 7 (Fig. 1). However participants were shown a geometric shape with a pre-trained associated value rather than pictures of money. Feedback was given as before, so participants were provided with a repeated reminder of the associated values throughout the experiment. Trials were split into blocks of low values – 10, 20 and 30p - and high values – 30, 40 and 50p. Each block was presented for half of each scanner run with a length of 60 trials per block. Thus there were 120 trials in each scanner run and two scanner runs separated by anatomical image collection. Block order for low and high values was counterbalanced between participants, with half experiencing HLHL and LHLH for the others. Payment following the scanning task was calculated by taking the outcome of a random subset of trials which would add up to £15. The adaptive staircase was not used in this experiment. The pre-training session was used to measure the length of response window that corresponded to 66% accuracy for each participant. Response windows were set to a length corresponding to 66% accuracy for each participant so that the average payout would be £10.

In order for participants to familiarise themselves with the task prior to the scanner and to learn the values of the Conditioned Stimuli used in the experiment, they completed a behavioural training session 4-7 days before scanning. In this session participants performed the same MID task they would complete in the scanner. This behavioural training was also used to set each participant's reaction time threshold so they would achieve 66% accuracy during the scanning task. During training the duration of the response window was controlled by an adaptive staircase, starting at 220ms then reducing by 1.5% for every two successful trials in a row and increasing by 1.5% following an unsuccessful trial.

5.3.2.3. Scanning Parameters

Scanning was performed in a 3T Phillips Achieva scanner with 8 channel phased array head coil. The same double echo, echo planar imaging sequence was employed as in the previous experiment. Each functional scanner run lasted 12 minutes and 15 seconds, collecting a total of 294 volumes of 36 slices for each TE using a voxel size of 3x3x3mm, TR = 2.5s, TE_1 = 20ms, TE_2 = 45ms.

5.3.2.4. Functional Data Analysis

To ensure accurate combination of data from both echoes, realignment parameters were calculated based on images collected with the first echo sequence and were then applied identically to image sequences for both echoes. As before, a weighted summation was then calculated combining both echoes into one time series upon which all subsequent processing and analysis was performed. Weighted images were then normalised to MNI space using participants' anatomical scans, before being smoothed with a 5mm FWHM Gaussian kernel.

The onsets of each CS formed the six regressors of interest (L10, L20, L30, H30, H40, H50) which were entered into the first level GLM. One regressor was entered representing button presses controlling for basic action preparation. A further twelve were entered representing win and lose feedback events of each value, controlling for effects in response to payout. Pairwise comparisons were calculated within subjects before being entered into second level analyses. Responding was compared for the effect of absolute highest vs lowest value: H50 vs. L10 and the effect of within block value: (H50+L30) vs. (H30+L10) to find regions which showed basic value dependent responding. Pure context effects independent of absolute value were also tested for by examining the overlap value: H30 vs. L30. A whole-brain uncorrected voxelwise threshold of $p < 0.001$ and an extent threshold of 31 voxels was applied in order to find significant activation clusters. This cluster size was calculated using AplhaSim (Cox, 1996) such that the corrected $\alpha = 0.05$. Once regions responding to value had been defined beta weights were extracted for all conditions so that it was possible to show the specific pattern of responding. Processing was performed using SPM8 and beta weights were extracted using marsbar (Brett, Anton, Valbregue and Poline, 2002). Weighted summation of echo signals was performed using code developed by the Sir Peter Mansfield Magnetic Resonance Centre.

5.3.3. Results and Discussion

5.3.3.1. Behavioural Results

The average response window for 66% success calculated during the training task was 234ms (S.D. 24ms) and it ranged from 272ms to 202ms. Accuracy during the scanning task was between 46.3% and 84.6% with a mean of 64.6% (S.D. 13.0%). This resulted in payments between £7.20 and £13.00 with an average of £9.85 (S.D. £2.06). As the range in performance was quite notable, a Pearson's correlation was conducted comparing winnings with the length of each participant's response window to ensure the manner in which this window was calculated was not causing the seemingly large variance in accuracy, and no such relationship was found ($r = 0.17$, $p > 0.05$). A six level, one-way ANOVA found no significant effect of trial value on reaction times ($F(1,5) = 1.11$) and a paired samples t-test showed no effect of block ($t(11) = 0.07$).

5.3.3.2. fMRI Results

5.3.3.2.1. Highest vs Lowest: H50 vs L10

In order to verify that the experiment had elicited value dependent responding, the highest and lowest absolute values were compared: H50 > L10 (Table 5.7). In line with previous studies, there is bilateral activation with its peak in the OFC which extends up into the ventral striatum and another bilateral activation in the Thalamus (Figure 5.5). There is also a strong effect in the Medial Frontal Gyrus (MFG) which suggests stimulus value modulates motor preparation even when there is no response choice to be made. There were no significant effects in the reverse contrast.

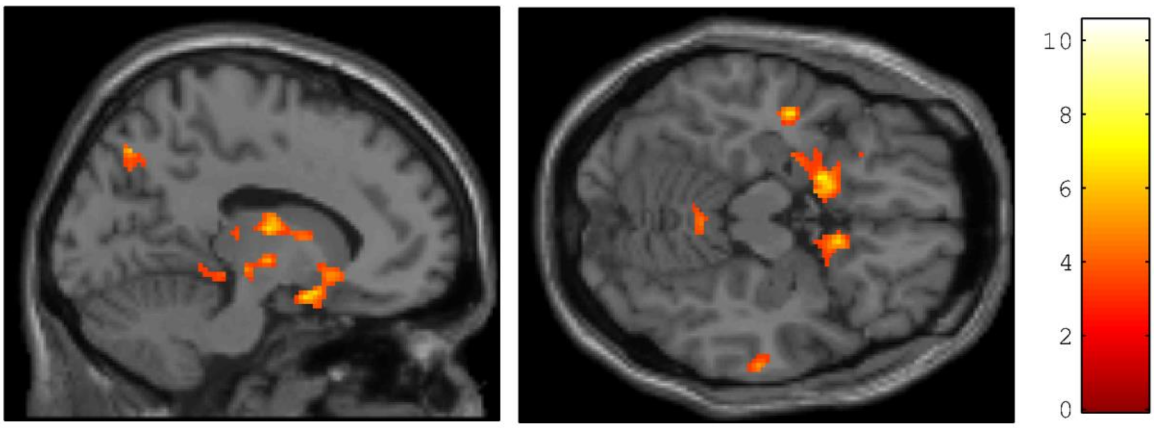

Figure 5.5 Significant clusters for Highest > Lowest Value: H50>L10

Table 5.7. Significant Clusters for, Highest > Lowest Value: H50>L10

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Medial Frontal Gyrus (L/R)	4	14	50	38	10.5
Precentral Gyrus (L)	-44	-8	48	46	8.7
Medial Frontal Gyrus (R)	6	4	54	31	8.6
Fusiform Gyrus (L)	-36	-28	-26	37	7.4
Orbito-Frontal Cortex/ Caudate Head (R)	12	14	-14	31	7.4
Orbito- Frontal Cortex/Caudate Head (L)	-16	6	-16	82	6.8
	-12	16	-8		4.7
Vental Lateral Nucleus (L)	-12	-10	12	49	6.4
Supramarginal Gyrus (R)	56	-38	24	41	6.1
	50	-32	30		5.8
	60	-36	34		4.4
Thalamus/Caudate Head (R)	8	0	10	45	6.0

5.3.3.2.2. Context Dependent Responding: (H50+L30) vs (H30+L10) and L30 vs H30

As it was hypothesized that value responding would be context dependent responses were also tested for regions displaying higher responding to high vs low values within blocks i.e. (H50+L30) > (H30+L10). Interestingly OFC activation is not apparent in this comparison, but there is still significant MFG activation, albeit at a weaker level than in Experiment 7 where pictures of cash elicited a stronger response (Table 5.8). There is also activation in the cingulate, although this is more posterior than is usually reported in value judgment experiments (Figure 5.6). There were no significant activations in the reverse contrast.

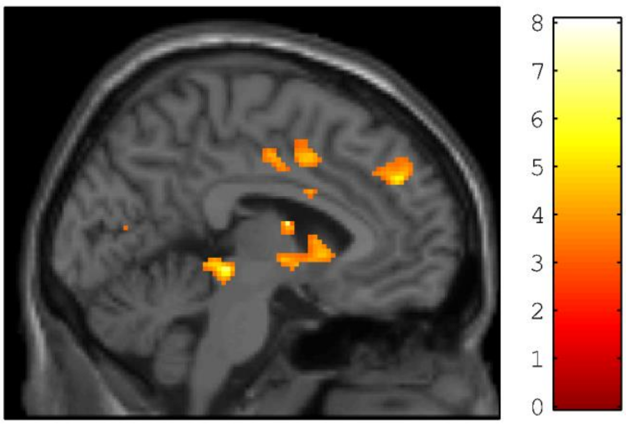

Figure 5.6 Significant Clusters for, Within Block Comparison: (H50+L30) > (H30+L10)

Table 5.8 Significant Clusters for, Within Block Comparison: (H50+L30) > (H30+L10)

Anatomical Region	Peak Activation (MNI coordinates)			Cluster Size (voxels)	Peak t-value
	X	Y	Z		
Cingulate (L/R)	-12	8	46	126	8.0
	4	0	46		7.4
Brain Stem	-2	-30	-6	70	7.7
Precentral Gyrus (R)	60	-2	30	57	6.8
Precentral Gyrus/Rolandic Operculum (R)	62	2	8	38	6.3
Superior Temporal Gyrus (L)	-50	-38	0	71	6.0
MFG (L)	-6	48	34	35	5.8
	-8	38	36		4.7
Cuneus (L)	-20	-76	-2	44	5.6

Although one would expect context dependent responding to be evident in the within block comparison, many other patterns of context independent responding would also be apparent. To identify only those regions that are completely re-scaling their responding patterns with changes in context, responding to the overlap value was compared between blocks: L30 > H30. Crucially, there were no areas of significant activation nor were there any in the reverse contrast: H30 > L30.

Beta weights for each area were then extracted and plotted so that overall patterns of responding could be assessed. Areas which respond to value do so either in a linear correlation with absolute value or in a manner which suggests higher responding to all values in high value blocks and lower responding to all values in low blocks (Figure 5.7).

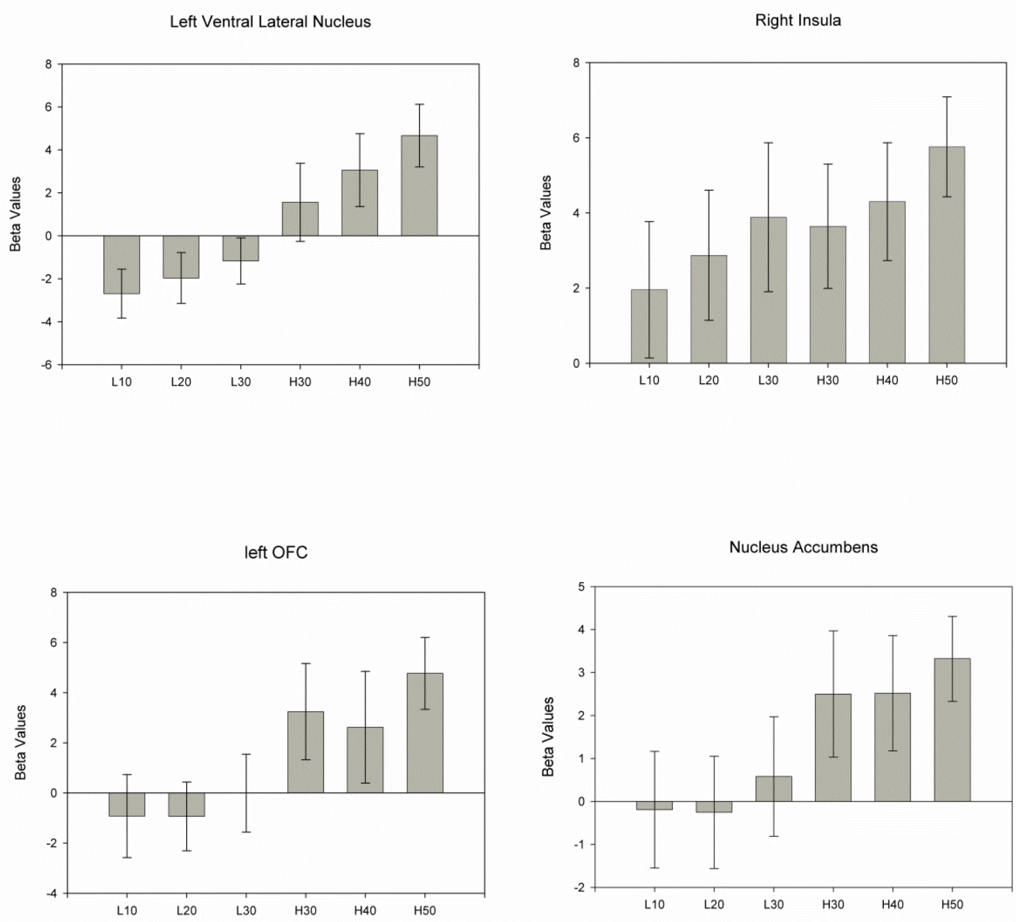

Figure 5.7 Extracted beta weights for the latter blocks.

5.3.4. Discussion

This experiment employed the same task and similar methodology as Experiment 7. The crucial difference was in the stimuli used: In this experiment all categorical cues were removed from the task. More similar values were used in both blocks, reducing the possibility of a qualitative difference between pence and pounds. These values were then paired with neutral geometric shapes and participants underwent a pre-training task so they were familiar with these values. This removes the visual cues of notes vs coins. Thus, the only plausible explanation for effects of context could be the temporal recency with which the other values were experienced. The results demonstrate that there is no longer an effect of block, therefore recency of experience (at least on the scale examined in this task) is not sufficient to create contextual effects. The recall bias for recent experiences does not generalise from memory experiments to valuation as one may expect. This has significant implications for DbS and future modelling of it and explains why efforts to apply an Ebbinghaus curve to the model in previous chapters have failed.

One potential criticism could be that as participants had been pre-trained to learn the associated values of stimuli, they began the task already calibrated to the full range of values they were going to see. Furthermore, they were aware of the manner in which blocks would cycle and that all stimuli would be seen multiple times, thus arguably negating the need to re-scale. However, neither of these suggestions can explain the results of Experiment 7 where participants did not know the full range of values at the start of the experiment. Both the VS and thalamus showed context dependent responding not only in the first two blocks but in the 3rd and 4th blocks. By that point in the experiment they were aware of the range of values. Therefore, the contextual effect would not be evident in the latter blocks.

The findings of Experiments 7 & 8 demonstrate a significant effect of rank ordering within context within the brain. They also show that recency of experience (at least on a shorter time scale) is not sufficient for these effects to occur. The next experiments address whether the effects of rank ordering can be replicated behaviourally and whether recency is necessary for such effects.

5.4. Experiment 9

5.4.1. Introduction

Experiments 7 & 8 suggest that the effect of rank ordering may well be stronger when the stimuli and task are simpler. Birnbaum (1992) showed individuals a simple gamble and asked them to estimate its value (or certainty equivalent). The potential answers which could then be selected had either a positive skew or a negative skew. When choosing from a negatively skewed response set participants were more likely to select higher values than when choosing from the positively skewed set. This was also observed by Stewart, Chater, Stott, & Reimers (2003), who manipulated the range and skew of the available responses showing that the rank order of options had a significant effect upon subsequent choice. Although consistent with DbS, findings relating to available response options could also be due to demand characteristics. Participants often avoid using the extremes of a response scale and many use the range of potential options as a cue towards what may be the “correct” answer (Kamenica, 2008; Prelec, Wernerfelt, & Zettelmeyer, 1997; Wernerfelt, 1995).

In addition to the effects of response set, DbS requires that the effects have a memory component and are not simply a result of menu context. Other experiments have demonstrated that recently viewed values can cause preference reversals (Stewart, 2009; Ungemach et al., 2011). Individuals were exposed to amounts that were either in between or outside of the payout values immediately prior to a critical question. This served to respectively increase or decrease the difference between values’ relative rank within recent experiences. However, this technique also serves to increase and decrease the overall range. Thus it is not possible to know whether the effect is driven purely by rank.

In a recent study Stewart et al., (in press) addressed both the issues of non-menu context and range effects. Participants answered a series of dilemmas based upon simple gambles. For half the participants the payout values had a significant positive skew whereas the remaining half experienced values with a negative skew. These two distributions result in very different predicted utility curves when calculated by DbS (Figure 5.8). When participants’ utility curves were extracted from their responses, there was a significant difference in the curvature of utility curves depending upon condition. The positive skew condition revealed a standard concave utility curve but participants in the negative skew condition actually exhibited a convex utility curve.

This experiment adapts the methodology used in Stewart et al (in press) in order to test the hypothesis suggested by the results of Experiments 7 & 8: Is context defined only by categorical cues rather than mere recency of exposure? If so, then it should be possible to replicate the results using a within subject design and concurrent exposure to different contexts.

5.4.2. Method

5.4.2.1. Participants

Fifty undergraduates at the University of Nottingham participated in the study for course credit, 6 males and 44 females. Their mean age was 18.27.

5.4.2.2. Stimuli

For the positive skew condition the values £10, £20, £50, £100, £200 and £500 were crossed with probabilities .2, .4, .6, .8 and 1 to create 300 items. All non-dominant pairings of these items were selected i.e., pairs where the higher value was not also paired with a higher probability. This was then repeated for the negative skew condition where values were created by subtracting the positive skew values from £510 meaning £10, £310, £410, £460, £490 and £500. These stimuli values were selected because of their significant positive and negative skew. This means that DbS predicts a concave and a convex utility curve respectively (Figure 5.8).

There were 150 non-dominant pairs for each condition. An additional 15 dominant pairs were also selected from each to serve as catch trials. Thus, participants made a total of 330 choices.

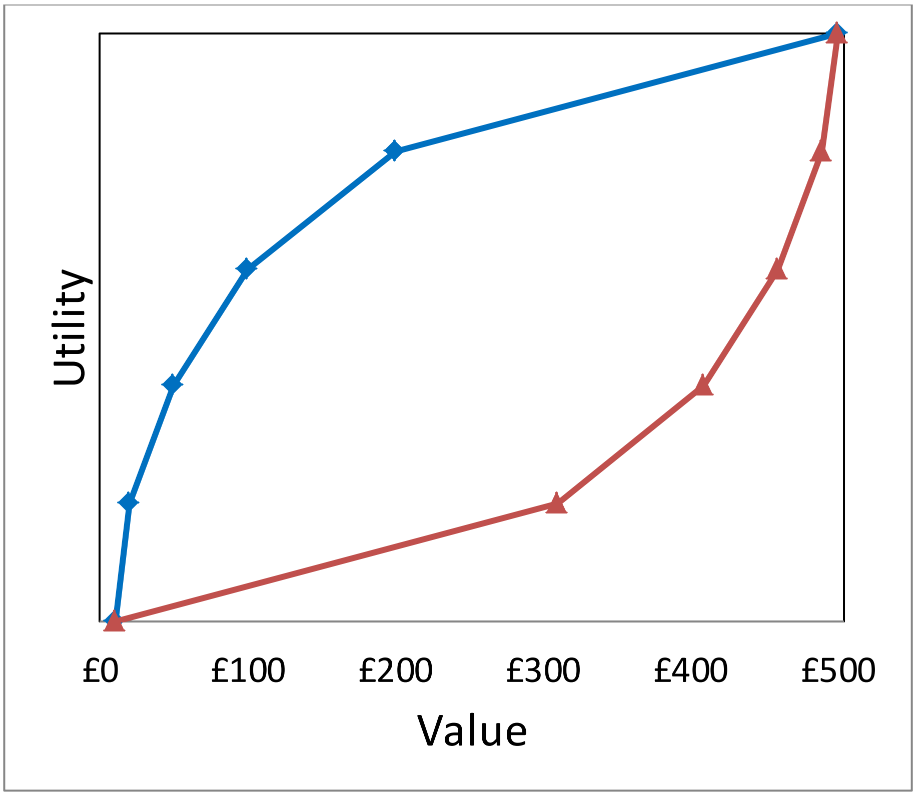

Figure 5.8 The utility curves predicted by DbS in positive and negative skew conditions

5.4.2.3. Procedure

The study was conducted online using open source software "Limesurvey". Participants were recruited through the University of Nottingham's online participant pool. When they signed up for the study they were then given a link which took them straight to the start of the experiment. They were then asked to complete the questions without breaks. They were also warned that although most questions had no right or wrong answer, there were catch questions which would be used to ensure they were paying proper attention. They were also informed that although they were free to complete the task at their own pace, if they provided no responses for a particularly long time the system would assume they had withdrawn and would time out.

Participants were told they would be shown potential choices from two hypothetical gameshows. One game-show offered prizes of phones, the other offered adventure days. In each gameshow two prizes of differing value would be offered, each linked to urns with black and gold balls in. The contestant would know how many of the balls were winning gold balls and how many were losing black ones. The contestant must decide whether to pick from the urn with the more valuable prize but usually fewer gold balls or from the urn with the less valuable prize but with a greater chance of drawing a gold ball. Choices were phrased as “A 20% chance of winning a phone worth £500 or a 100% chance of winning a phone worth £10” or in the alternate gameshow “A 60% chance of winning an adventure day worth £310 or an 80% chance of winning an adventure day worth £490”.

Questions were presented with five on each web-page and the gameshow in question alternated from one page to the next. Thus, participants would answer five questions from the positive skew condition, then five from the negative skew condition. The positive and negative skews were counterbalanced between gameshows, with half of participants seeing phones with positively skewed values and the other half seeing adventure days with positively skewed value. To make the current game-show as salient as possible pictures of either mobile phones or people on adventure days were shown at the top of each page. The background colour on either side of the questions was also alternated between red and black depending upon the current game-show.

5.4.3. Results

The probability of selecting the left option on any given choice can be estimated using Equation 1. This equation uses the subjective utility (U) of each individual payout of X for risky gambles and Y for safe gambles. These are then weighted by the probabilities Q and P for risky and safe gambles respectively so that the Luce decision rule can be used to calculate the probability of selecting the safe option. The γ component controls the determinism of the equation, with values above 1 making the resulting predictions more confident for smaller differences in expected utilities.

$$Prob(Safe) = \frac{bias_{cond}[q.u(y)]^{\gamma_{cond}}}{bias_{cond}[q.u(y)]^{\gamma_{cond}} + [p.u(x)]^{\gamma_{cond}}}$$

By performing a log transform this can then be re-written as

$$log\left[\frac{prob(safe)}{prob(safe)}\right] = log(bias_{cond}) + \gamma_{cond}\log\big(u(y)\big) - \gamma_{cond}\log\big(u(x)\big) + \gamma_{cond}\log\left(\frac{q}{p}\right)$$

One can then preform the substitutions

$$\log(bias_{cond}) = \beta_{cond} + cond.\beta_{cond}$$

And

$$\gamma_{cond} = \beta_{log\frac{q}{p}} + cond.\beta_{cond.log\left(\frac{q}{p}\right)}$$

Setting cond as a dummy variable representing experimental condition (positive or negative skew dilemma) gives

$$log\left[\frac{prob(safe)}{1 - prob(safe)}\right] = \beta_0 + \beta_{cond} + \sum_i \beta_i.X_i + \beta_{log\left(\frac{q}{p}\right)}.log\left(\frac{q}{p}\right) + \beta_{cond.log\left(\frac{q}{p}\right)}cond.log\left(\frac{q}{p}\right)$$

This is now a standard logarithmic regression equation and can be analysed with standard statistical packages. To simplify each term point by point, β_0 is the overall bias towards choosing the safe option. β_{cond} is the dummy variable indicating the bias towards selecting safe in one condition over the other. The term $\sum_i \beta_i.X_i$ is a series of i dummy variables, one for each value used in the experiment. The dummy indicates the presence of each value in either gamble, 1 for the safe option, -1 for risky. The influence of the relative difference in probability between the two options is represented by $\beta_{log\left(\frac{q}{p}\right)}.log\left(\frac{q}{p}\right)$ and the difference between conditions represented by $\beta_{cond.log\left(\frac{q}{p}\right)}cond.log\left(\frac{q}{p}\right)$.

To extract the term u(X) from the beta weights the determinism had to be first controlled and then the exponent calculated to transform back into the original scale.

$$u(X) = exp\left(\frac{\beta_i}{\gamma_{cond}}\right)$$

The analysis was applied to the results and the extracted utility curves are plotted in Figure 5.9. There is a clear visual difference between the lines with mid-range payouts being weighted lower in the negative skew condition. This is confirmed using a two-sample z-test comparing utilities of £310 in the negative skew condition and £200 in the positive skew condition. Despite the former having a much higher objective value its estimated utility is significantly lower ($Z = 9.65$, $p<0.001$).

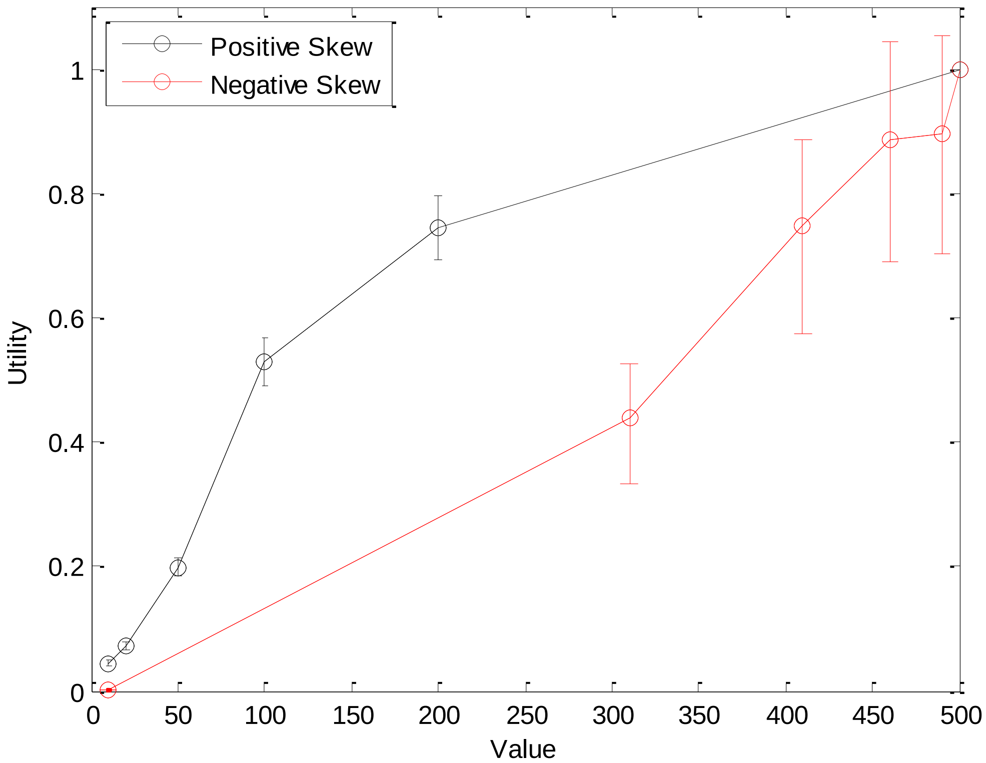

Figure 5.9 Weights for probability (left) and Utility of payouts (right). Error bars represent standard errors of parameter estimates.

5.4.4. Discussion

This experiment shows that the neuroimaging findings presented in Experiments 7 & 8 have demonstrable behavioural correlates. Previous studies have shown that the distribution of experiences can manipulate individuals' utility curves. However, this is the first demonstration using a within-subject design with concurrent exposure to different distributions. The design used here also controls for effects of range, as both distributions have the same maximum and minimum. Furthermore, as the differences in distribution are only apparent across several dilemmas, menu effects can also be ruled out along with their potential demand effects.

5.5. Chapter Discussion

This chapter presents strong support for rank order encoding of value using both neuroimaging and behavioural methods. Experiments 7 & 8 show that regions of the brain which are responsive to reward encode value by rank order. There was no region in which activity as measured by fMRI represented the true financial value of the current trial. The findings also show that value is encoded in a distributed network with varying levels of context dependency. Regions which are associated with lower level encoding such as the Thalamus and VS show a context dependency that is defined by categorical cues. However, regions such as the vmPFC and the ACC that are further along the processing chain and aggregate information from multiple inputs are unaffected by these cues. Instead, they encode value in terms of the item's overall rank within the experiment. Experiment 9 used choices between simple gambles in different distributions and demonstrated a strong effect of rank with utility curves modified by the environment distribution.

Both Experiment 8 & 9 demonstrate that the boundaries of contextual effects are defined by explicit cues of category membership. Experiment 8 demonstrates that simple recency of experience is not sufficient to cause contextual effects, whilst Experiment 9 shows that it is not necessary either. This finding explains why attempts in previous chapters to add forgetting functions such as the serial order position curve and Ebbinghaus curve were unsuccessful. It is likely that such recency effects will be apparent over very long durations, but this would be due to complete forgetting or inability to recall previous experiences. The effect here is attributable only to items being explicitly disregarded when they were still available to be sampled.

Experiment 9 also demonstrates a successful extension of neuroeconomic findings to novel behavioural effects. This is a troublingly rare occurrence. Many findings within neuroscience have not been demonstrated in a behavioural analogue and it seems there is a serious lack of cross modality replication. This chapter not only provides novel findings relevant to important and current topics of discussion but also presents a strong bridge between the fields of neuroscience and behavioural psychology.

6. Chapter 6 – General Discussion

6.1. Introduction

Recent years have seen a growing interest in describing human decision making and judgements using psychologically plausible process models. Progress towards this goal has speeded as more psychologists become interested in financial decision making and more economists become interested in results from psychology. This thesis adds to the debate and provides novel empirical findings relevant to the development of more accurate and plausible models. The main focus of the empirical research has been the Decision by Sampling model, but many of the results also speak to more general issues and debates within the JDM community. This chapter will begin with a summary of the empirical findings from each chapter before bringing together the findings from all experiments and discussing the implications for wider issues.

6.1.1. Summary of Findings

The overarching findings are that in complex multi-attribute decisions DbS is a relatively poor model of human judgement and decision making, but in more simple tasks, with small numbers of stimuli values and few attributes, DbS performs well. Both behavioural and neuroimaging findings strongly support rank order encoding and context dependency in simple decision environments.

Chapter 2 employed a multi-attribute valuation task and uses participants' estimates of apartment valuations to test the explanatory power of DbS by explicitly simulating the model. The distribution of item values is also modified in a manner that tests the predictions of RFT. The results show that DbS is a relatively poor predictor of value estimates, being outperformed by a simple baseline measure and a Weighted ADDitive model. Implementing a weighting function to simulate effects of recency in memory sampling did not improve the performance of DbS and parameter estimation revealed that the best fit was from a model closest to the original unweighted DbS. The results were more supportive of RFT, but only when one assumes the full range of values was underestimated due to the higher values being comparatively rare in one condition.

Chapter 3 again used a multi-attribute valuation task with apartments as the stimuli, but this time the distributions of individual attribute values were modified rather than the overall values of the items. Therefore, DbS made specific predictions about the way these scales would be interpreted and weighted in participants' estimates of overall item value. The results showed that participants' use of these attributes did not match the predictions of DbS, but nor did they match the predictions of other considered models. As the results could be explained by assumptions and information garnered from real world experience prior to the experiment, the same stimuli were used in Experiment 4 but with a different cover story. When participants believed they were judging the value of mineral deposits instead of apartments, their use of attributes changed significantly. The change suggested that the results of Experiment 3 were attributable to prior expectations. However, the results for valuations of mineral deposits were also not explained by any considered model. The pattern of responding suggested that the majority of variance could be explained by WADD. But, for attributes participants considered important, values near the extremes of the scale were relatively over-weighted suggesting they were especially salient and used as a qualitative cue to raise or lower estimates. The predictions of DbS were less accurate than simple baseline measures in both of the experiments.

Chapter 4 used similar multi-attribute stimuli to Experiment 3, but employed a choice task rather than eliciting value estimates. DbS again performed relatively poorly. When modelled separately for each participant, DbS was always outperformed by either a WADD model, Dawes rule or both. However, there was some tentative support for encoding by rank ordering as when participants' use of attribute values was recovered, the weighting function represented cumulative frequency. Experiment 6 also used concurrent eye tracking to examine whether participants visual attention to different attributes was a reliable analogue for the differential weighting they subsequently applied to information when making their decisions. No such correlation was found. As a result, when attention weighted drift diffusion models were fitted to the data their accuracy was poor. This suggests that eye-tracking cannot serve as a meaningful analogue for attention terms in models such as DbS and Decision Field Theory (Busemeyer & Townsend, 1993).

Chapter 5 focussed more closely upon the neuroscience evidence for rank encoding and context sensitivity. Experiment 7 used an fMRI paradigm with a distribution of values which meant rank order encoding predicted a different activation pattern than models assuming a simple linear transform of absolute value. The results showed that activity in value sensitive regions was highly context dependent and that response magnitudes were predicted better by rank order encoding than they were by absolute value. Experiment 8 then demonstrated that the context sensitivity found in dopaminergic reward regions including the VS and the thalamus were only sensitive to contextual shifts when there was an exogenous cue. That is to say that these regions were not disproportionately sensitive to the most recent events. Experiment 9 then sought to demonstrate these contextual effects and a basic effect of DbS by eliciting choices between risky gambles in different contexts. Each context presented values drawn from either a positively or negatively skewed distribution and participants' utility curves showed significant concavity and convexity in these respective conditions. This is precisely the pattern predicted by DbS.

6.2. Judgement and Decision by Rank-Order

The behavioural results presented in the preceding chapters are mixed in their support of encoding by rank order. First let's examine the results which support DbS and rank order effects before moving on to discuss those which do not. Experiment 6 shows that in a multi-attribute choice task when one of the attributes is given a non-linear distribution, the rank ordered cumulative frequency plot can be recovered from participants decisions using multiple regression. Put simply, participants are using rank order not the attribute's absolute value. Furthermore, Experiment 9 demonstrates that participants' utility curves are not stable over time but respond to the distribution of values in the choice set. When making decisions in a context with a positive skew the recovered utility curves exhibit the concave shape familiar from EUT and CPT. However, when making decisions in a context with significant negative skew the utility curves exhibit a convex shape, suggesting risk seeking behaviour.

Although Experiment 6 found that participants' use of the crime attribute was best explained by rank ordering, the results of valuation tasks in Experiments 3&4 were very different. For both experiments, one attribute had a cubic distribution such that rank ordering would predict a cubic effect in participants' use of the attribute when calculating their responses. In Experiment 3, a cubic effect was present, but the curvature was in the opposite direction to that predicted. Furthermore, the same effect was observed for an attribute with linear distribution, but curvilinear weighting upon true values. In Experiment 4 there was no cubic effect for the attribute with a cubic distribution, but there was for two others despite them having linear distributions. The difference between these results and those of the choice task show that participants make different use of information depending upon the type of response elicited.

There is a significant literature investigating the difference between choice and valuation (Hsee et al., 1999; Lichtenstein & Slovic, 1971; Sevdalis & Harvey, 2006; Tversky et al., 1990). The results of Experiment 5 concur with the evaluability hypothesis, arguably the most widely supported explanation for the phenomenon. It is found that although the large majority of participants use the same WADD process for judgements and decisions, their interpretation and use of the attribute values differs between task modalities. When providing value estimates, participants relatively under-weight crime (the attribute with the cubic distribution) as the majority of its values cluster around the centre of the scale. This makes it harder to discriminate relatively small absolute differences when only one item is seen at a time. However, the attribute's influence increases significantly during the choice task as presenting two side by side makes differences more apparent. Furthermore, individuals are more likely to recall items and values more similar to those presented because they act as a cue (Brown et al., 2007; Howard & Kahana, 2002). In the valuation task there will be no particular effect of this bias. However, in choice the two items will cue recall of other similar values. As values are clustered around the centre of the scale the close proximity of many items to the cues means there will be a higher number of recalled values which lie between the two alternatives.

Although there is no evidence for rank ordering of attribute values during valuation tasks, Experiment 2 does show evidence for range-frequency effects. RFT can provide a parsimonious explanation of the results if one makes the reasonable assumption that participants are prone to underestimating the upper bound when values at the extreme high end of a range are particularly rare. Several other accounts can explain why participants regularly provide value estimates that fall within the unrepresented portion of the value range, including a simple regression to the mean. However, RFT is the only theory which also predicts the control items' rise in value estimates in the modified distribution condition. Thus, in valuation tasks the interpretation of attribute scales is not reliant upon rank order, but rank does have a significant impact upon value estimates.

A variable which is undoubtedly a factor in the size of rank order effects is the complexity of the choice/judgement environments. The strongest behavioural effect is found for Experiment 9, the experiment with the simplest task and design. The only attributes are payout and probability, and there are only 6 possible payouts in each condition. This means that it is possible for – the majority of – participants to represent all potential payouts in working memory concurrently. Indeed, the vast majority of previous demonstrations of rank order effects have used either a manipulation with no memory requirement i.e. menu context, (Birnbaum, 1992) or distributions with a small number of values (Stewart, 2009; Stewart et al., 2005). This issue is discussed in more detail below.

6.3. Multiple Attribute Tasks and Information Integration

As stated above, the reported experiments demonstrate that the complexity of the task has a significant impact upon how closely participants' choices match the predictions of DbS. Therefore the obvious question is whether DbS can only make accurate predictions in simple decision environments. If true, this would suggest that participants switch to a qualitatively different strategy in more complex environments, in a manner similar to the adaptive toolbox account (Bröder, 2003; Gigerenzer & Selten, 2002). Alternatively it may be that participants do not use different strategies, but specific characteristics of DbS are incorrect in its current formulation. If these inaccuracies were to have a minimal effect upon predictions in simpler environments but a more significant effect in more complex environments then one would expect the results found throughout this thesis.

One of the components of complexity is the number of different values experienced for an attribute within a choice environment. When the task environment is simple and there are relatively few values which are repeated frequently throughout the task, DbS is far more accurate at predicting decisions. Another source of complexity is the number of attributes for which values are given. Again, in Experiment 9, there are only two attributes: value and probability. However, in Experiments 1-6, items have between 4 and 5 attributes. One effect of this additional complexity is to make the task more cognitively demanding through the retention, recall and examination of more information. It also raises the more specific question of how individuals judge the importance of each attribute and then combine very different kinds of information when calculating a valuation or decision for the item as a whole. Thus there are two general factors which may explain why the accuracy of DbS is sensitive to complexity: General cognitive load (especially the number of values which can potentially be sampled from memory) and the number of attributes to be considered.

First consider the effect of general cognitive load and the number of values experienced on a single attribute scale during the task. DbS predicts choices most accurately in Experiment 9 where there are relatively few values experienced during the experiment. Accuracy is far lower in Experiments 5&6 where a large number of different values are sampled from a continuous scale with little repetition of values. However, Experiment 6 also shows that participants' use of values on a single attribute is modified by rank order. This suggests that DbS correctly predicts aspects of the early stages in the decision process. So the later steps must account for a large proportion of the drop in accuracy. Principally, the way DbS integrates information from different attributes.

In its original form DbS predicts that individuals sample from each attribute with equal probability. This essentially equates to an un-weighted additive model, where it is the rank positions which are being summed rather than the attribute values. In every experiment where this was examined here, the accuracy was significantly improved by allowing the weighting of attributes to vary freely within the model. This weighted formulation of DbS matches the predictive accuracy of WADD, but does not outperform it. This includes Experiments 5&6 where the non-linear attribute distribution should favour DbS, given the evidence for rank ordering.

This is a curious result. One suggestion could be that the integration of information is multiplicative, not additive. However it is unlikely that this is driving the effect, as simulations show that multiplicative and additive models can reliably mimic each other (Stewart, 2011) . Another possibility is that participants are integrating information in a way which violates the independence of value perception on each attribute. For example it seems plausible that participants could think a particular floor space dimension is very good for a one bedroom apartment, but poor when there are four bedrooms. But this cannot be so easily argued for Experiment 4 where participants were instead valuing mineral deposits. Furthermore there is no reason to suspect that such violations of independence would have a disproportionate effect upon weighted DbS over WADD as they both predict the same mechanism for information integration.

DbS and WADD have the same explanatory power even in experiments where their predictions are less highly correlated, which suggests that the models explain independent variance. This could be the result of over-fitting, as the independent weighting of each attribute adds a large number of free parameters to the models. If this is the case, then the results still show tentative support for WADD, as both have the same number of free parameters but more individuals are categorized as responding using that strategy than DbS. However, this would not explain the rank ordering of attribute values.

It is possible that participants use a variety of cues and arguments when making complex decisions. In Experiments 5&6, Dawes rule was shown to modulate the performance of other models, but was a poor predictor when considered alone. It was theoretically possible for DbS to explain this phenomenon with a recency weighting curve, negating any suggestion of separate cues or processes. However, adding such a weighting curve reduced DbS's accuracy. Thus it seems that participants use a variety of cues/strategies including WADD, rank order and Dawes when making their decisions. This appears an unsatisfactory suggestion as it seems particularly complex and suggests a high level of processing if all are considered for each decision. However it seems more plausible that on any particular choice an individual may rely primarily on WADD. Then if they are relatively indifferent between options, the fact that one particular attribute value was better than a large number of previous items, or one of the current options is better on more attributes may be used as a reason or cue for choosing the favoured item. In essence, the results are compatible with a form of the adaptive toolbox (Bröder, 2003; Gigerenzer & Selten, 2002), but using tools or models which are more complex than proposed by the original heuristic approach.

6.3.1. Attention as a Measure of Attribute Weighting

Preferential looking toward the chosen item is a critical prediction of attention driven models such as decision field theory and several versions of drift diffusion (Busemeyer & Townsend, 1993; Krajbich et al., 2010). The gaze cascade supports these predictions, allowing such models to predict choices from eye movements with accuracy significantly better than chance (Atalay et al., 2012; Glaholt & Reingold, 2009; Shimojo et al., 2003; Simion & Shimojo, 2007). However, the models also make specific predictions of a correlation between attention towards information and subsequent information weighting in choice. Previous eye-tracking experiments which have found results in favour of such attention weighted models have used single attribute items or have used paradigms in which the eye-tracking data did not discern which attributes were attended (Atalay et al., 2012; Glaholt & Reingold, 2009; Krajbich et al., 2010; Ratcliff & McKoon, 2007). This is a serious weakness, because this essentially confounds attention towards the item with attention towards information. Experiment 6 is the first study to orthogonalise item and information in a multi-attribute choice task and the results do not support these model's predictions. There is no correlation between relative attention to attributes and their relative influence upon responses. Furthermore, when visual attention is used to weight a drift diffusion model (Krajbich et al., 2010), the resulting predictions are significantly less accurate than other baseline measures.

One response to this negative finding could be to retain the existing models and their assumptions regarding attention, but to dismiss visual attention as a reliable analogue for individuals' covert attention. However the gaze cascade effect is known to be a robust phenomenon (Shimojo et al., 2003; Simion & Shimojo, 2007) and relative visual attention between items does predict 65% of choices in Experiment 6. Therefore, to say that eye-tracking cannot predict any aspects of the decision process is to ignore these results and merely leave a different phenomenon without an explanation. What these results do suggest is that the feedback loop hypothesised in experiments on the gaze cascade (but for an opposing view see Orquin & Mueller Loose, 2013) is only sensitive to the item's overall value. For the attention and saccade planning systems to be biased towards the preferred item in the lead up to a choice, suggests that there is a correlation between attention planning and the current relative evidence, or accumulator values for the items. It also suggests that the value of each item is represented as a whole at all time-points, rather than as separate accumulators for separate attributes and the information being integrated only when the choice is made. So the eye tracking results support a model of evidence accumulation for options, with value only being represented at the whole item level. But, there is no effect of attention to different information or attributes. Evidence accumulator models therefore need to be re-examined with respect to information weighting, as none currently make satisfactory predictions.

6.4. Sampling and Memory Effects

One of the main aims of this thesis was to investigate the impact of memory phenomena upon decision making. Many of the characteristics of recall and sampling are unspecified in DbS, which is surprising given how important memory is to the model. The two undefined characteristics investigated in this thesis are how many items are sampled and which items are most likely to be recalled.

6.4.1. Sample Size

Previous experiments have suggested that judgements and decisions rely upon comparisons with a sample no larger than an individual's WM capacity (Dougherty & Hunter, 2003a, 2003b; Kareev et al., 1997; Sprenger & Dougherty, 2006). However, DbS predicts a running total of accumulated evidence is represented for each item while alternatives are sampled sequentially. Therefore, DbS does not predict that all sampled alternatives must be concurrently represented in WM, so the number of sampled alternatives is potentially larger. The results of Experiment 9 show that when the number of previously experienced values is small enough to fit in WM, DbS accurately predicts choices. However, when the number of previous values is very large, such as in Experiments 1-6, the model has poor predictive accuracy and explicitly estimating parameters shows that the best model fit comes from the largest possible sample.

Experiments 1-6 used a larger number of values, meaning the best fitting sample size could only be estimated by explicitly modelling DbS. This means that any general inaccuracies in the DbS model would reduce the accuracy and validity of the estimated parameters. As other analyses have demonstrated that DbS is not accurate, particularly in multi-attribute tasks, it is perhaps not surprising that attempts at modelling the sample size found that the best fit came with the largest possible sample, which also resulted in the strongest correlation between DbS and WADD. However, the results of multi-attribute tasks in Experiments 1-6 show participants have far greater discriminability than would be possible using only the 7 +/-2 comparisons, or possibly fewer (Cowan, 2001), that could be concurrently represented in WM. Such a small number of comparisons would also result in the majority of attributes receiving only one comparison with alternative values. This would mean the weighting functions extracted for individual attributes would be far noisier than those which were found.

Despite Experiments 1-6 rejecting very small samples, Experiment 9 shows reliable rank order and context effects when the relatively small number of alternative values can be represented simultaneously in WM. This can be interpreted as tentative support for the suggestion that participants switch strategies depending upon the number of values from which they can sample (Lindskog et al., 2013). However, having fewer potentially sampled values also makes it far easier to infer which are likely to be present in any single sample. This in turn reduces the noise in any modelling and estimation. This noise and uncertainty is an inherent problem when modelling a stochastic system such as covert memory recall, which cannot be directly measured. Thus in Experiments 1-6 it's possible that participants are sampling from memory (drawing a sample larger than WM) but this is not detected because simulations of the model must average across a large number of potential samples. For a direct test of sample sizes, future experiments must be designed with a specific manipulation which can test for effects of sample size. Many previous experiments have done so, but by comparing the response patterns of individuals with high WM capacity to those with low capacity (Dougherty & Hunter, 2003a; Kareev et al., 1997) it is possible that the differences are caused by a more general underlying difference in cognitive ability or strategy at the individual level, rather than a within-individual switch from one strategy to another (Gaissmaier et al., 2006).

6.4.2. Predicting Which Items are Sampled

To date the research on DbS has focussed primarily upon the distribution of all values within the environment. As detailed in chapter one, this has found a strong correlation between the global distributions of values and average weighting or utility curves (Stewart, 2009; Stewart et al., 2006). However, this is only one mechanism which can systematically bias judgements based upon DbS.

Equally important is which values are sampled at each particular time point. If there is a systematic bias to recalling say, particularly high or low value items, then this will bias DbS judgements to be lower or higher respectively. This is true regardless of the overall distribution of values from which individuals' sample. To examine the likely patterns of such biases, evidence and models were sought from the literature on memory research.

Perhaps the most robust finding in all memory research is the recency effect. The likelihood of an item being successfully recalled is inversely proportional to the length of time which has passed since it was encoded (Ebbinghaus, 1913). Therefore it was hypothesised that values experienced most recently were more likely to be sampled and thus bias valuations and decisions. However, when DbS was modelled with either an Ebbinghaus or serial order position curve used to weight the probability of previous values being sampled, no improvement was found. The weighted model never surpassed the fit of the original instantiation of DbS. When the effect of different parameter values are examined the fit of the weighted model shows an asymptotic relationship with the performance of the unweighted model. The accuracy of the modified model approached that of the original DbS model as changes in the Ebbinghaus curve parameter reduced the strength of the relative recency bias.

Other studies of value estimation have shown that the single most recent item value has a significant influence upon the next valuation (Matthews & Stewart, 2009). But this is essentially a demonstration of the anchoring effect (Simmons, LeBoeuf, & Nelson, 2010; Tversky & Kahneman, 1974). It is also not a specific prediction of DbS, which is primarily intended to predict choices. If participants had a steep forgetting curve then the model would predict an anchoring effect from individual attribute values, but not from the overall item value. Although this pattern of insufficient adjustment has been found on individual attribute scales in relatively simple decisions (Ungemach et al., 2011), it is not found here in more complex multi-attribute tasks. Furthermore, these previous studies merely demonstrate the effect of the single most recent experience. There was no attempt to show a continuous forgetting curve or an effect comparable to models of memory.

In the experiments reported in this thesis there was only one situation where the results showed significant overweighting of recent binary comparisons. This was in chapter 4 where Dawes rule was show to add explanatory power orthogonal to both WADD and DbS, even when DbS was weighted with an Ebbinghaus curve. But this bias only occurs for the immediately available alternative, i.e. when there is no memory requirement. Therefore it cannot be argued as an effect of recency as there is no such bias for the immediately preceding items.

Another reliable finding in memory research is that when one item has been provided as a cue, or is simply recalled first, then individuals are more likely to recall other items which are more similar to it (Brown et al., 2007; Howard & Kahana, 2002) and more likely to recall other items experienced in the same context (Godden & Baddeley, 1980). In the context of DbS this seems logical: there is little point in an individual recalling the price of a chocolate bar they purchased last week when judging the relative merits of two differently priced cars. Experiment 9 demonstrates that this memory phenomenon is apparent in participants' decision making. When participants alternated between choosing phones and adventure days of different values, their inferred utility curves closely matched the distribution of values within each item category. This creation of salient contextual boundaries had a strong effect, with the neuroscience evidence also showing the same pattern. There was no effect of recency or block, but a strong contextual distinction in neural responses when an exogenous context cue is present. This contextual finding is predicted by models of memory, however it is also compatible with more general models which suggest that the use of strategies and meta-information is specific to choice environments and tasks (Hertwig et al., 2006).

6.4.3. Exemplar or Non-Exemplar Representation

The findings that the best fitting sample size is one so large that it encompasses all values in the experiment and that there is no effect of recency seriously questions the role of memory in decision making. It speaks to a debate more general than simply testing the specifics of DbS: Are experienced distributions represented in an exemplar or non-exemplar system (Camilleri & Newell, 2013)? DbS clearly falls into the former category, as it predicts that specific instances and exemplars are recalled from memory for each decision. There are also other more general models which posit exemplar representation (Lejarraga, Dutt, & Gonzalez, 2012).

Non-exemplar models suggest that rather than remembering or recalling the specific instances of previous items, values or experiences, individuals instead store meta-data. Depending upon the particular formulation of the model, this can include the mean, range, distribution and skew of the experienced values (Brainerd & Reyna, 1990; Hertwig et al., 2006; Kühberger, 1998). It is generally suggested that these models do not predict any recency bias (Camilleri & Newell, 2013) as specific items are not being recalled. Furthermore, as the meta-information is updated with each new experience, it makes predictions similar to that of an exemplar model with no limit on sample size. This means non-exemplar models can explain the lack of memory phenomena in Experiments 1-6. And, as different meta-information would be stored for different choice environments, the context effects of Experiments 7 & 9 would still be predicted. Such a model could also explain the cubic weighting functions found in Chapter 4, if one posits that the meta-information represents (or is at least biased by) rank order or range-frequency effects.

It would therefore seem that non-exemplar models are a more suitable hypothesis. However, they are by no means a perfect solution. For one, if the meta-information is updated with every new experience, it seems likely that there would still be a bias towards the most recent information. It would be a very difficult task to perfectly calculate the change which should be made to an estimated distribution based upon the importance of the new information and the importance of all previous information.

When examining the results of previous experiments, the balance of findings also support an exemplar based model (Camilleri & Newell, 2013). For example, in recent model competitions for decisions from experience, exemplar models have reliably outperformed non-exemplar alternatives (Erev et al., 2010; Gonzalez & Dutt, 2011; Hau, Pleskac, Kiefer, & Hertwig, 2008). Exemplar models also provide a more parsimonious explanation for idealised distributions of experiences producing more accurate judgements (Giguere & Love, 2013) and individuals acting as though the distribution of experienced examples is the true distribution, even when they have explicit knowledge that it is not (Feiler et al., 2013).

What is most interesting with regards the results presented here are the effects of WM capacity. Non-exemplar models cannot explain the oversensitivity to correlations (Kareev et al., 1997) and reduced accuracy in specific judgements (Dougherty & Hunter, 2003a, 2003b) found for individuals with smaller WM. The results of Experiments 7 & 9 showed that the exemplar model of DbS predicts the data very accurately when there are few values or items. However, in Experiments 1-6 when there were more values than could be represented in WM, the exemplar model performed poorly and the results fit well with a non-exemplar account. Therefore, the results presented here suggest that when the number of values within a task or choice environment is low enough that they can all be represented concurrently in WM, participants rely upon the exemplar model of DbS. However, when there are too many values and items, individuals cannot rely upon such a strategy they instead use meta-information.

6.5. Neural Encoding of Value

Experiment 7 is the first fMRI study to directly compare the predictions of rank order and absolute financial value representations in neural encoding. The results are strongly in favour of rank order encoding in higher functioning regions including the vmPFC and the ACC. These are thought to be closest to the end of the information integration process and therefore the nearest neurological proxy for the final response (Chib, Rangel, Shimojo, & O'Doherty, 2009; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). The response of lower, dopaminergic reward regions is also compatible with DbS, as responding rescales to represent the rank order of the current item within the current context. This total re-scaling is incompatible with other models of relative difference from the average environment value, which would predict smaller activations when the differences from the average value were also smaller (Knutson & Cooper, 2005).

Early fMRI experiments were limited by the quality of scanners and analysis techniques. The poor signal to noise ratio and high cost of testing large numbers of participants meant the most successful and popular paradigms were those which relied upon relatively simple comparisons (O'Doherty et al., 2001). These involved participants receiving gains and experiencing losses of large or small magnitudes, sometimes with varying degrees of risk attached. When greater activity was found for higher values or more preferable outcomes the simplest assumption was that these areas responded with more activation to more valuable stimuli and opportunities. Thus there has been an implicit assumption running through subsequent research that there is a linear relationship between value and neural response.

As the accuracy of fMRI has improved, research has moved towards addressing more complex questions. These include the endowment effect (Plassmann, O'Doherty, & Rangel, 2007), menu effects in choice (Cooper et al., 2009; Elliott et al., 2008) and the interaction between visual attention and value signal (Lim et al., 2011); but the most basic assumptions and findings have not been re-visited. The original assumption of a linear relationship has persisted, not because there was strong evidence to support it, but because it was the simplest interpretations and there was no evidence to directly contradict it. Many researchers have also been reluctant to examine more basic characteristics when the field has moved on to relatively complex issues. Furthermore, despite the advances in scanners, researchers still tend to prefer simpler designs with qualitative comparisons and contrasts, in an attempt to ensure high power and minimise the chances of null-effects in what is still a very expensive methodology. It is thanks largely to the advances in scanner technology and to the new double echo procedures being pioneered at the University of Nottingham that it was possible to conduct the experiments reported in Chapter 5 and re-examine this core building block of neuroeconomics.

Previous papers have claimed to find evidence of absolute encoding (Knutson & Bossaerts, 2007; Kringelbach & Rolls, 2004) and even specific patterns of prospect theory such as loss aversion (Nieuwenhuis et al., 2005; Tom, Fox, Trepel, & Poldrack, 2007). However, these studies have all contained characteristics which mean that DbS and rank ordering still predicts their findings. The experiments claiming to demonstrate an effect of loss aversion have not found disproportionately stronger activation to losses than to gains of identical value, as one might have assumed. Instead, they began from the assumption that loss aversion would exist and therefore deliberately biased the stimuli values to counteract the predicted effect. The studies used gains which were twice as large as the concomitant losses (Tom et al., 2007) or two thirds larger (Nieuwenhuis et al., 2005). When the same degree of deactivation and activation was found for losses and gains respectively, it was interpreted as evidence for loss aversion and prospect theory within the brain. However, because the gains and losses had the same ranks within their distributions, DbS also predicts that they will elicit the same strength of neural signal. The results of Experiment 7 are relevant not only to these two particular experiments, but to all studies in neuroeconomics which rely upon the assumption of continuous value representation. Many experiments must now be re-evaluated and a significant proportion will subsequently be open to different interpretation.

Chapter 5 also demonstrates a novel finding with regards context. In Experiment 7 the mid-brain reward regions of the thalamus and ventral striatum responded not to the value's rank order within the entire experiment, but to its rank within the current block. Possibly the most obvi ous suggestion if one is to take inspiration from the memory literature, is that the simple recency of experiences in the current block causes the effect. However, if this were the case then the same effect would also have been found in the VS during experiment 8, when the same blocks and timings were used but other exogenous contextual cues were removed. Furthermore, previous reports of context dependency have all used some form of exogenous cue such as the colour of a card (Nieuwenhuis et al., 2005), buying vs. selling (De Martino, Kumaran, Holt, & Dolan, 2009; Knutson et al., 2008), whether the value is part of a gamble or not (Cooper et al., 2009) or the visual shape of the stimulus (Kringelbach & Rolls, 2004; O'Doherty et al., 2001).Thus Chapter 5 demonstrates that these exogenous cues are necessary to create the context dependent encoding found in Experiment 7.

The behavioural results of Experiment 9 also show that exogenous cues cause contextual dependency without any temporal component, demonstrating significantly different utility curves in different contexts. In this experiment the exogenous cues were present in the form of different visual cues (background colour) and cover story (gameshow for phones or adventure days), but questions were presented intermixed. Thus, the effect of recency and block was eliminated, and yet the effect remained, showing that these exogenous cues are both necessary and sufficient.

One issue that is not addressed by these experiments is which specific exogenous cues cause context effects. Considering all available evidence, there are several possibilities: obviousness of block switch, cover story, visual cues, and the categorizability of values themselves. Each of these possible explanations will now be addressed in turn.

The obviousness of the switch from one block or context to another may be a factor as there was a gap of ~20s between blocks in experiment 7 which was not present in Experiment 8. There was also a similar short gap between blocks in Cooper et al. (2009). However, given the lack of a general effect of recency and the successful findings of context dependency within a single scanner block (Elliott et al., 2008; Knutson & Cooper, 2005) it is certainly not necessary for creating context effects, and is unlikely to be sufficient.

The cover story given for the different contexts seems an intuitively strong manipulation, and this characteristic is present in several studies including Experiment 9 (Cooper et al., 2009; De

Martino et al., 2009). The only study where this could arguably be the sole manipulation is De Martino et al. (2009) which relies upon the endowment effect. However, interpreting endowment as merely a cover story is troublesome and likely too simplistic given the evidence suggesting that endowment changes behaviour and perceptions of value (Horowitz & McConnell, 2002; Isoni, Loomes, & Sugden, 2011). Therefore it is not currently possible to isolate the effect of a change in cover story using solely neuroscience evidence. If behavioural evidence is considered then the results of Experiments 3&4 suggest that the cover story is indeed sufficient. This is possibly only true when it has a direct influence upon individuals' interpretation of the values themselves, as in Experiment 9, where such possibilities are (largely) removed, there are still visual cues.

Visual cues are seemingly ubiquitous among demonstrations of context dependency. The only experiments which do not have an explicit visual cue are Experiment 8, which finds no effect, and Cooper et al. (2009). Including the latter is also debatable as one could easily argue that having a probability present next to the value, acts as a visual cue. Therefore it seems that the presence of a visual cue is likely sufficient and possibly necessary, but the null hypothesis has not been explicitly tested.

Categorizability of values themselves is not an issue often addressed in the literature. However, it is potentially a very strong manipulation. In some experiments its presence is obvious. For example, in Experiment 7 the values in different blocks differed by an order of magnitude and the ranges did not overlap. In others it is not as intuitively obvious but still exists: In Experiment 9 one has a situation where the positive skew values were all relatively round numbers (£10, £20, £50, £100 etc.), whereas those in the negative skew were not (£310, £410, £480, £490 etc.). In fact there is evidence that this difference alone is strong enough to cause the contextual effects observed in Experiment 9, even when no other contextual cue or manipulation is applied (Stewart and Reimers, unpublished manuscript). Other tasks have used numbers which are similar in magnitude, but as the context is that of gains vs. losses it is the sign attached to the value which provides its categorization.

In summary, the available evidence suggests that the cover story and obviousness of the switch from one block to another may well be sufficient to cause context dependency, but the evidence is equivocal. Furthermore, there is enough evidence to say that neither characteristic is necessary. For visual cues and the categorizability of the values themselves, there is good evidence to say that they are sufficient to cause context dependency and there is also not currently enough evidence to rule them out as necessary characteristics.

6.6. Conclusions and Summary

The primary aim of this thesis was to examine Decision by Sampling as a model of human choice and value judgement. The results show a clear dichotomy, with the model accurately predicting responses in simple tasks and in environments with few values, but performing poorly in more complex task environments. Rank order effects are found in Experiments 7-9, exactly matching the predictions of DbS. For each of these tasks, the experiment or decision context had a maximum of 6 possible values, this means that all values could be represented in working memory at the same time, making the relative ranks more salient. However, in the more complex task environments of Experiments 1-6 the large number of values which could be potentially recalled means DbS must rely upon models of memory to predict which items will enter the sample. In this case, not only does DbS perform no better than a standard WADD model, but the results do not replicate the most universal findings of memory research: recently experienced items are sampled no more frequently than those very temporally distant. The fact that the core memory sampling assumption of the model cannot replicate such a common effect, strongly suggests that specific examples are not being recalled from memory. Thus the data suggest a switch in behaviour, depending upon the properties of the task: Participants use DbS and rank ordering when all relevant values can be concurrently held in WM, but rely on a qualitatively different process when the number of values in the choice environment exceeds the capacity of WM.

One suggestion in the literature is that individuals instead hold meta-information about the distribution (Brainerd & Reyna, 1990). If one assumes that meta-information and utility curves are built upon the relative frequencies of experienced values then they would have the same global rank based predictions of DbS, but there is no memory component and specific examples are not recalled. There is some evidence suggesting that this is not the case and that explicit sampling takes place (Feiler et al., 2013; Kareev et al., 1997; Lindskog et al., 2013). However, it should be noted that the majority of this evidence comes from tasks which are again very simple and use a small number of values, or involve responses which cannot be calculated using only distribution data. Furthermore, Experiments 3-6 show that participants' responses are not based upon stable underlying meta-data or estimates of the distribution. They are manipulated by both response mode and the cover story assigned to the task.

The eye tracking evidence of Experiment 6 supports evidence accumulation over time. The characteristics of the gaze cascade show that participants' attention becomes more biased towards the preferred item in the last moments before a decision is made. This increase correlates with the hypothesised accumulation of evidence for the preferred item over time and the relative differences in the underlying item-wise accumulator values predicted by DbS and other evidence accumulation models. These models are also supported by the increased reaction times for choices where there was little relative difference between options (Glöckner & Betsch, 2008; Glöckner & Herbold, 2011). However, none of the evidence accumulation models considered here could explain participants' weighting of attributes in light of the visual attention given to each.

Experiment 6 is the first to examine the prediction common to several models, of attribute weighting by relative attention. The results strongly suggest that previous findings have been driven by a gaze cascade towards a preferred item. This means that existing eye tracking evidence needs to be fundamentally reconsidered as it does not separate the effects of item value and attribute weighting in measures of attention. Future research must first attempt to replicate this finding across other types of decisions. If the result proves robust, then a new model must be developed. The most promising is an evidence accumulation model, as this will still explain the gaze cascade and reaction time effects. But, one which accumulates evidence based upon noisy weighted addition of the available information, as this best explains the behavioural results.

Future work will need to examine more closely the effect of environment complexity upon decisions. Research should investigate whether there is indeed a number of values at which individuals show a qualitative shift in their strategies. Based upon the existing evidence, this is likely to be at a number equal to or slightly higher than the capacity of an individual's WM. Further to that line of research, the question then becomes what strategy individuals switch to. The best prediction based upon the behavioural data appears to be a standard WADD model. Furthermore, the eye tracking results suggest a noisy accumulator with evidence accumulation rates based upon a WADD integration of information, so the obvious starting point for future research is drift diffusion (Ratcliff, 2001; Ratcliff & McKoon, 2007).

7. References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science, 46*(11), 1497-1512.

Abdellaoui, M., Barrios, C., & Wakker, P. P. (2007). Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory. *Journal of Econometrics, 138*(1), 356-378.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396-408.

Ariely, D., & Wallsten, T. S. (1995). Seeking subjective dominance in multidimensional space: An explanation of the asymmetric dominance effect. *Organizational Behavior and Human Decision Processes, 63*(3), 223-232.

Atalay, A. S., Bodur, H. O., & Rasolofoarison, D. (2012). Shining in the Center: Central Gaze Cascade Effect on Product Choice. *Journal of Consumer Research, 39*(4), 848-866. doi: 10.1086/665984

Ayal, S., & Hochman, G. (2009). Ignorance or integration: the cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making, 22*(4), 455-474. doi: 10.1002/bdm.642

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0908104107

Beckstead, J. W. (2008). Modeling sequential context effects in judgment analysis: A time series approach. *Judgment and Decision Making, 3*(7), 570-584.

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

Birnbaum, M. H. (1992). Violations of monotonicity and contextual effects in choice-based certainty equivalents. *Psychological Science, 3*(5), 310-314.

Birnbaum, M. H., & LaCroix, A. R. (2008). Dimension integration: Testing models without trade-offs. *Organizational Behavior and Human Decision Processes, 105*(1), 122-133. doi: 10.1016/j.obhdp.2007.07.002

Biswas, D., Grewal, D., & Roggeveen, A. (2010). How the order of sampled experiential products affects choice. *Journal of Marketing Research, 47*(3), 508-519.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*(4), 700.

Boyce, C. J., Brown, G. D. A., & Moore, S. C. (2010). Money and Happiness: Rank of Income, Not Income, Affects Life Satisfaction. *Psychological Science, 21*(4), 471-475. doi: 10.1177/0956797610362671

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity Perception: Perceptual Anchor Model of Context-Coding. *Journal of the Acoustical Society of America, 76*(3), 722-731.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*(1), 3-47. doi: http://dx.doi.org/10.1016/0273-2297(90)90003-M

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113*(2), 409-432. doi: 10.1037/0033-295x.113.2.409

Brehmer, B. (1980). In One Word - Not From Experience. *Acta Psychologica, 45*(1-3), 223-241. doi: 10.1016/0001-6918(80)90034-7

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*(2), 137-154.

Bröder, A. (2002). Take The Best, Dawes' Rule, and Compensatory Decision Strategies: A Regression-based Classification Method. *Quality & Quantity, 36*(3), 219-238. doi: 10.1023/a:1016080517126

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(4), 611-625. doi: 10.1037/0278-7393.29.4.611

Brown, G. D. A., Gardner, J., Oswald, A. J., & Qian, J. (2008). Does wage rank affect employees' well-being? *Industrial Relations, 47*(3), 355-389. doi: 10.1111/j.1468-232X.2008.00525.x

Brown, G. D. A., & Matthews, W. J. (2011). Decision by Sampling and Memory Distinctiveness: Range Effects from Rank-Based Models of Judgment and Choice. *Frontiers in Psychology, 2*.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*(3), 539.

Brunswick. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 2*(63), 24.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation, 32*, 275-318.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive Modeling*: SAGE Publications.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432.

Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience, and the continuum of uncertainty in risky choice. *Decision making: Neural and behavioural approaches, Progress in brain research. Oxford, UK: Elsevier*.

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience, 29*(39), 12315-12320. doi: 10.1523/jneurosci.2575-09.2009

Cohen, J. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*: L. Erlbaum Associates.

Cook, R. L., & Stewart, T. R. (1975). A comparison of seven methods for obtaining subjective descriptions of judgmental policy. *Organizational Behavior and Human Performance, 13*(1), 31-45. doi: http://dx.doi.org/10.1016/0030-5073(75)90004-5

Cooper, J. C., Hollon, N. G., Wimmer, G. E., & Knutson, B. (2009). Available alternative incentives modulate anticipatory nucleus accumbens activation. *Social Cognitive and Affective Neuroscience, 4*(4), 409-416. doi: 10.1093/scan/nsp031

Cosmides, L., & Tooby, J. (1994). Better than rational: Evolutionary psychology and the invisible hand. *The American Economic Review, 84*(2), 327-332.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162-173. doi: 10.1006/cbmr.1996.0014

Dawes, R. M. (1979). Robust Beauty of Improper Linear Models in Decision Making. *American Psychologist, 34*(7), 571-582. doi: 10.1037//0003-066x.34.7.571

De Martino, B., Kumaran, D., Holt, B., & Dolan, R. J. (2009). The Neurobiology of Reference-Dependent Value Computation. *Journal of Neuroscience, 29*(12), 3833-3842. doi: 10.1523/jneurosci.4832-08.2009

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 23*(4), 968-986. doi: 10.1037//0278-7393.23.4.968

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science, 14*(2), 175-180.
Dieckmann, A., Dippold, K., & Dietrich, H. (2009). Compensatory versus noncompensatory models for predicting consumer preferences. *Judgment and Decision Making, 4*(3), 200-213.
Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica, 113*(3), 263-282. doi: http://dx.doi.org/10.1016/S0001-6918(03)00033-7
Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*(6), 968-982.
Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1352.
Ebbinghaus. (1913). Memory: A Contribution to Experimental Psychology.
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*(9), 2213-2218. doi: 10.1111/j.1460-9568.2008.06202.x
Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23*(1), 15-47.
Feiler, D. C., Tong, J. D., & Larrick, R. P. (2013). Biased Judgment in Censored Environments. *Management Science, 59*(3), 573-591. doi: 10.1287/mnsc.1120.1612
Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 966.
Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social Sampling Explains Apparent Biases in Judgments of Social Environments. *Psychological Science, 23*(12), 1515-1523. doi: 10.1177/0956797612445313
Garner, W. R. (1954). Context effects and the validity of loudness scales. *Journal of Experimental Psychology, 48*(3), 218.
Gigerenzer, G., & Goldstein, G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650-669. doi: 10.1037/0033-295x.103.4.650
Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*: MIT Press.
Gigerenzer, G., & Todd, P. M. (2000). *Simple heuristics that make us smart*: Oxford University Press, USA.
Giguere, G., & Love, B. C. (2013). Limits in decision making arise from limits in memory retrieval. *Proceedings of the National Academy of Sciences of the United States of America, 110*(19), 7613-7618. doi: 10.1073/pnas.1219674110
Glaholt, M., & Reingold, E. (2009). Stimulus exposure and gaze bias: A further test of the gaze cascade model. *Attention, Perception, & Psychophysics, 71*(3), 445-450. doi: 10.3758/app.71.3.445
Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., . . . Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 355.
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. Bonn: Max Planck Institute for Research on Collective Goods.
Glöckner, A., & Herbold, A.-K. (2011). An Eye-tracking Study on Information Processing in Risky Decisions: Evidence for Compensatory Strategies Based on Automatic Processes. *Journal of Behavioral Decision Making, 24*(1), 71-98. doi: 10.1002/bdm.684
Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition, 123*(1), 21-32. doi: http://dx.doi.org/10.1016/j.cognition.2011.12.002

Godden, D., & Baddeley, A. (1980). When does context influence recognition memory? *British Journal of Psychology, 71*(1), 99-104.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118*(4), 523.

Gould, J. D. (1973). Eye movements during visual search and memory search. *Journal of Experimental Psychology, 98*(1), 184.

Gowland, P. A., & Bowtell, R. (2007). Theoretical optimization of multi-echo fMRI data acquisition. *Physics in Medicine and Biology, 52*(7), 1801.

Hadland, K. A., Rushworth, M. F. S., Gaffan, D., & Passingham, R. E. (2003). The Anterior Cingulate and Reward-Guided Selection of Actions. *Journal of Neurophysiology, 89*(2), 1161-1164. doi: 10.1152/jn.00634.2002

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience, 28*(22), 5623-5630. doi: 10.1523/jneurosci.1309-08.2008

Harvey, N., & Fischer, I. (2005). Development of experience-based judgment and decision making: The role of outcome feedback. *The routines of decision making*, 119-137.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21*(5), 493-518.

Henderson, J. M., Weeks Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*(1), 210-228. doi: 10.1037/0096-1523.25.1.210

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2006). The role of information sampling in risky choice. *Information sampling and adaptive cognition*, 72-91.

Hey, J. D. (2001). Does repetition improve consistency? *Experimental economics, 4*(1), 5-54.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*(1), 1-55.

Horowitz, J. K., & McConnell, K. E. (2002). A Review of WTA/WTP Studies. *Journal of Environmental Economics and Management, 44*(3), 426-447. doi: http://dx.doi.org/10.1006/jeem.2001.1215

Howard, M. W., & Kahana, M. J. (2002). When Does Semantic Similarity Help Episodic Retrieval? *Journal of Memory and Language, 46*(1), 85-98. doi: http://dx.doi.org/10.1006/jmla.2001.2798

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576-590. doi: 10.1037//0033-2909.125.5.576

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 90-98.

Hunt, L. T., Kolling, N., Soltani, A., Woolrich, M. W., Rushworth, M. F. S., & Behrens, T. E. J. (2012). Mechanisms underlying cortical activity during value-guided choice. *Nat Neurosci, 15*(3), 470-476. doi: http://www.nature.com/neuro/journal/v15/n3/abs/nn.3017.html#supplementary-information

Isoni, A., Loomes, G., & Sugden, R. (2011). The Willingness to Pay—Willingness to Accept Gap, the “Endowment Effect,” Subject Misconceptions, and Experimental Procedures for Eliciting Valuations: Comment. *The American Economic Review, 101*(2), 991-1011. doi: 10.1257/aer.101.2.991

Johnson, E. J., Haubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology-Learning Memory and Cognition, 33*(3), 461-474. doi: 10.1037/0278-7393.33.3.461

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*(1), 263-272. doi: 10.1037/0033-295x.115.1.263

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: a naive sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*(4), 441-480.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454.

Kahneman, D., & Tversky, A. (1979). Prospect Theory - Analysis of Decision Under Risk. *Econometrica, 47*(2), 263-291.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111*(4), 1072-1098.

Kamenica, E. (2008). Contextual inference in markets: On the informational content of product lines. *The American Economic Review, 98*(5), 2127-2149.

Kareev, Y. (2003). On the perception of consistency. *Psychology of Learning and Motivation, 44*, 261-285.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*(3), 278.

Knutson, B., & Bossaerts, P. (2007). Neural antecedents of financial decisions. *Journal of Neuroscience, 27*(31), 8174-8177. doi: 10.1523/jneurosci.1564-07.2007

Knutson, B., & Cooper, J. C. (2005). Functional magnetic resonance imaging of reward prediction. *Current Opinion in Neurology, 18*(4), 411-417.

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., & Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *Neuroreport, 12*(17), 3683-3687.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience, 25*(19), 4806-4812. doi: 10.1523/jneurosci.0642-05.2005

Knutson, B., & Wimmer, G. E. (2007). Splitting the difference - How does the brain code reward episodes? *Reward and Decision Making in Corticobasal Ganglia Networks, 1104*, 54-69. doi: 10.1196/annals.1390.020

Knutson, B., Wimmer, G. E., Rick, S., Hollon, N. G., Prelec, D., & Loewenstein, G. (2008). Neural antecedents of the endowment effect. *Neuron, 58*(5), 814-822. doi: 10.1016/j.neuron.2008.05.018

Kobayashi, S., Pinto de Carvalho, O., & Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *J Neurosci, 30*(2), 534-544. doi: 10.1523/jneurosci.4009-09.2010

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292-1298.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci, 12*(5), 535-540. doi: http://www.nature.com/neuro/journal/v12/n5/suppinfo/nn.2303_S1.html

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology, 72*(5), 341-372. doi: 10.1016/j.pneurobio.2004.03.006

Kühberger, A. (1998). The Influence of Framing on Risky Decisions: A Meta-analysis. *Organizational Behavior and Human Decision Processes, 75*(1), 23-55. doi: http://dx.doi.org/10.1006/obhd.1998.2781

Leh, S. E., Ptito, A., Chakravarty, M. M., & Strafella, A. P. (2007). Fronto-striatal connections in the human brain: A probabilistic diffusion tractography study. *Neuroscience Letters, 419*(2), 113-118. doi: 10.1016/j.neulet.2007.04.049

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based Learning: A General Model of Repeated Binary Choice. *Journal of Behavioral Decision Making, 25*(2), 143-153.
Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*(1), 46-55. doi: 10.1037/h0031207
Lim, S.-L., O'Doherty, J. P., & Rangel, A. (2011). The Decision Value Computations in the vmPFC and Striatum Use a Relative Value Code That is Guided by Visual Attention. *Journal of Neuroscience, 31*(37), 13214-13223. doi: 10.1523/jneurosci.1246-11.2011
Lindskog, M., Winman, A., & Juslin, P. (2013). Naive Point Estimation. *Journal of Experimental Psychology-Learning Memory and Cognition, 39*(3), 782-800. doi: 10.1037/a0029670
Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565-572. doi: 10.1037/0096-1523.4.4.565
Loomis, J. B. (1990). Comparative reliability of the dichotomous choice and open-ended contingent valuation techniques. *Journal of Environmental Economics and Management, 18*(1), 78-85. doi: http://dx.doi.org/10.1016/0095-0696(90)90053-2
Louie, K., & Glimcher, P. W. (2012). Efficient coding and the neural representation of value. *Ann N Y Acad Sci, 1251*(1), 13-32. doi: 10.1111/j.1749-6632.2012.06496.x
Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141*(2), 233.
Maltby, J., Wood, A. M., Vlaev, I., Taylor, M. J., & Brown, G. D. A. (2012). Contextual Effects on the Perceived Health Benefits of Exercise: The Exercise Rank Hypothesis. *Journal of Sport & Exercise Psychology, 34*(6), 828-841.
Mantonakis, A., Rodero, P., Lesschaeve, I., & Hastie, R. (2009). Order in Choice Effects of Serial Position on Preferences. *Psychological Science, 20*(11), 1309-1312.
Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior, 11*(5), 649-653.
Matthews, W. J. (2012). How Much Do Incidental Values Affect the Judgment of Time? *Psychological Science, 23*(11), 1432-1434. doi: 10.1177/0956797612441609
Matthews, W. J., & Stewart, N. (2009). Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks. *Judgment and Decision Making, 4*(1), 64-81.
Mellers, B. A. (1980). Configurality in multiple-cue probability learning. *The American Journal of Psychology*, 429-443.
Melrose, K. L., Brown, G. D. A., & Wood, A. M. (2013). Am I Abnormal? Relative Rank and Social Norm Effects in Judgments of Anxiety and Depression Symptom Severity. *Journal of Behavioral Decision Making, 26*(2), 174-184. doi: 10.1002/bdm.1754
Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for value-based choice response times under high and low time pressure. *Judgement & Decision Making, 5*, 437-449.
Mori, S., & Ward, L. M. (1995). Pure Feedback Effects in Absolute Identification. *Perception & Psychophysics, 57*(7), 1065-1079.
Mosteller, F., & Nogee, P. (1951). An Experimental Measurement of Utility. *Journal of Political Economy, 59*(5), 371-404. doi: 10.2307/1825254
Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*(1), 53-65. doi: 10.1037/0278-7393.29.1.53
Nieuwenhuis, S., Heslenfeld, D. J., von Geusau, N. J. A., Mars, R. B., Holroyd, C. B., & Yeung, N. (2005). Activity in human reward-sensitive brain areas is strongly context dependent. *Neuroimage, 25*(4), 1302-1309. doi: 10.1016/j.neuroimage.2004.12.043

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience, 4*(1), 95-102.

Olsson, A.-C., Enkvist, T., & Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(6), 1371.

Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica, 144*(1), 190-206. doi: http://dx.doi.org/10.1016/j.actpsy.2013.06.003

Pachur, T., Hertwig, R., & Steinmann, F. (2012). How Do People Judge Risks: Availability Heuristic, Affect Heuristic, or Both? *Journal of Experimental Psychology-Applied, 18*(3), 314-330. doi: 10.1037/a0028279

Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological Review, 72*(6), 407.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 16*(2), 366-387.

Philiastides, M. G., Biele, G., & Heekeren, H. R. (2010). A mechanistic account of value computation in the human brain. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1001732107

Philiastides, M. G., & Ratcliff, R. (2013). Influence of Branding on Preference-Based Decision Making. *Psychological Science*. doi: 10.1177/0956797612470701

Pitz, G. F., Heerboth, J., & Sachs, N. J. (1980). Assessing the Utility of Multiattribute Utilty Assessments. *Organizational Behavior and Human Performance, 26*(1), 65-80.

Plassmann, H., O'Doherty, J., & Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *Journal of Neuroscience, 27*(37), 9984-9988. doi: 10.1523/jneurosci.2131-07.2007

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*(6741), 233-238.

Poser, B. A., Versluis, M. J., Hoogduin, J. M., & Norris, D. G. (2006). BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. *Magnetic Resonance in Medicine, 55*(6), 1227-1235. doi: 10.1002/mrm.20900

Posse, S., Wiese, S., Gembris, D., Mathiak, K., Kessler, C., Grosse-Ruyken, M. L., . . . Kiselev, V. G. (1999). Enhancement of BOLD-contrast sensitivity by single-shot multi-echo functional MR imaging. *Magnetic Resonance in Medicine, 42*(1), 87-97.

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*(1), 118-126.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica, 68*(5), 1281-1292.

Rakow, T., Newell, B. R., & Zougkou, K. (2010). The role of working memory in information acquisition and decision making: Lessons from the binary prediction task. *The Quarterly Journal of Experimental Psychology, 63*(7), 1335-1360.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology, 20*(2), 262-270. doi: 10.1016/j.conb.2010.03.001

Rashotte, M. A., & Wedell, D. H. (2012). Context effects on tempo and pleasantness judgments for Beatles songs. *Attention Perception & Psychophysics, 74*(3), 575-599. doi: 10.3758/s13414-011-0255-y

Ratcliff, R. (2001). Putting noise into neurophysiological models of simple decision making. *Nature Neuroscience, 4*(4), 336-336.

Ratcliff, R., & McKoon, G. (2007). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation, 20*(4), 873-922. doi: 10.1162/neco.2008.12-06-420

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging, practice, and perceptual tasks: a diffusion model analysis. *Psychology and aging, 21*(2), 353.
Ratcliff, R., & Van Dongen, H. P. (2009). Sleep deprivation affects multiple distinct cognitive processes. *Psychonomic Bulletin & Review, 16*(4), 742-751.
Reilly, B. A., & Doherty, M. E. (1989). A note on the assessment of self-insight in judgment research. *Organizational Behavior and Human Decision Processes, 44*(1), 123-131. doi: http://dx.doi.org/10.1016/0749-5978(89)90038-1
Reilly, B. A., & Doherty, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes, 53*(3), 285-309. doi: http://dx.doi.org/10.1016/0749-5978(92)90067-H
Rettinger, D. A., & Hastie, R. (2001). Content effects on decision making. *Organizational Behavior and Human Decision Processes, 85*(2), 336-359.
Robinson, A., Loomes, G., & Jones-Lee, M. (2001). Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making, 21*(1), 17-27.
Rogers, R. D., Ramnani, N., Mackay, C., Wilson, J. L., Jezzard, P., Carter, C. S., & Smith, S. M. (2004). Distinct portions of anterior cingulate cortex and medial prefrontal cortex are activated by reward processing in separable phases of decision-making cognition. *Biological Psychiatry, 55*(6), 594-602. doi: 10.1016/j.biopsych.2003.11.012
Rushworth, Matthew F. S., Noonan, MaryAnn P., Boorman, Erie D., Walton, Mark E., & Behrens, Timothy E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. *Neuron, 70*(6), 1054-1069. doi: 10.1016/j.neuron.2011.05.014
Schoenbaum, G., Setlow, B., & Ramus, S. J. (2003). A systems approach to orbitofrontal cortex function: recordings in rat orbitofrontal cortex reveal interactions with different learning systems. *Behavioural Brain Research, 146*(1-2), 19-29. doi: 10.1016/j.bbr.2003.09.013
Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*(3), 610.
Serences, J. T. (2008). Value-Based Modulations in Human Visual Cortex. *Neuron, 60*(6), 1169-1181.
Sevdalis, N., & Harvey, N. (2006). Determinants of willingness to pay in separate and joint evaluations of options: Context matters. *Journal of Economic Psychology, 27*(3), 377-385. doi: 10.1016/j.joep.2005.07.001
Seymour, B., & McClure, S. M. (2008). Anchors, scales and the relative coding of value in the brain. *Current Opinion in Neurobiology, 18*(2), 173-178. doi: 10.1016/j.conb.2008.07.010
Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nat Neurosci, 6*(12), 1317-1322. doi: http://www.nature.com/neuro/journal/v6/n12/suppinfo/nn1150_S1.html
Simion, C., & Shimojo, S. (2007). Interrupting the cascade: Orienting contributes to decision making even in the absence of visual stimulation. *Perception & Psychophysics, 69*(4), 591-595. doi: 10.3758/bf03193916
Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology, 99*(6), 917.
Snook, B., Dhami, M. K., & Kavanagh, J. M. (2011). Simply criminal: Predicting burglars' occupancy decisions with a simple heuristic. *Law and Human Behavior, 35*(4), 316-326. doi: 10.1007/s10979-010-9238-0
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*(2), 202-211.
Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and aging, 25*(2), 377.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62*(6), 1041-1062. doi: 10.1080/17470210902747112

Stewart, N. (2011). Information Integration in Risky Choice. *Frontiers in Psychology, 2*. doi: 10.3389/fpsyg.2011.00301

Stewart, N., & Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 416.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categori zation of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 3.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1-26. doi: 10.1016/j.cogpsych.2005.10.003

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*(1), 23.

Sunstein, C. R., & Thaler, R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*: Yale University Press New Haven.

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307*(5715), 1642-1645. doi: 10.1126/science.1105370

Todd, F. J., & Hammond, K. R. (1965). Differential Feedback in 2 Multiple-Cue Probability-Learning Task. *Behavioral Science, 10*(4), 429-435. doi: 10.1002/bs.3830100406

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315*(5811), 515-518.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature, 398*(6729), 704-708.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352.

Tunney, R. J. (2006). Preference reversals are diminished when gambles are presented as relative frequencies. *The Quarterly Journal of Experimental Psychology, 59*(9), 1516-1523. doi: 10.1080/17470210600750509

Tunney, R. J., Mullett, T. L., Gardner, A., & Moross, C. (2012). Does the Butcher on the Bus Phenomenon Require a Dual Process Explanation? A Signal Detection Analysis. *Frontiers in Psychology, 3*. doi: 10.3389/fpsyg.2012.00208

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*(4), 281-299. doi: 10.1037/h0032955

Tversky, A., & Kahneman, D. (1974). Judgement Under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory - Cumulative Presentation of Uncertainty. *Journal of Risk and Uncertainty, 5*(4), 297-323. doi: 10.1007/bf00122574

Tversky, A., Slovic, P., & Kahneman, D. (1990). The Causes of Preference Reversal. *American Economic Review, 80*(1), 204-217.

Ungemach, C., Stewart, N., & Reimers, S. (2011). How Incidental Values From the Environment Affect Decisions About Money, Risk, and Delay. *Psychological Science, 22*(2), 253-260. doi: 10.1177/0956797610396225

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review, 108*(3), 550.

Vlaev, I., & Chater, N. (2007). Context effects in games: Local versus global sequential effects on choice in the prisoner's dilemma game. *Judgment and Decision Making, 2*(5), 380-389.

Vlaev, I., Chater, N., & Stewart, N. (2007). Relativistic financial decisions: Context effects on retirement saving and investment risk preferences. *Judgment and Decision Making Journal, 2*(5), 292-311.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences, 15*(11), 546-554.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Watkinson, P., Wood, A. M., Lloyd, D. M., & Brown, G. D. A. (2013). Pain ratings reflect cognitive context: A range frequency model of pain perception. *Pain, 154*(5), 743-749. doi: 10.1016/j.pain.2013.01.016

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and Decision Making. *Annual Review of Psychology, 60*, 53-85. doi: 10.1146/annurev.psych.60.110707.163633

Wernerfelt, B. (1995). A rational reconstruction of the compromise effect: Using market data to infer utilities. *Journal of Consumer Research*, 627-633.

Westenberg, M. R. M., & Koele, P. (1994). Multi-attribute evaluation processes: Methodological and conceptual issues. *Acta Psychologica, 87*(2–3), 65-84. doi: 10.1016/0001-6918(94)90044-2

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 198.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*(6), 1411.

Windschitl, P. D., & Young, M. E. (2001). The influence of alternative outcomes on gut-level perceptions of certainty. *Organizational Behavior and Human Decision Processes, 85*(1), 109-134.

Wood, A. M., Boyce, C. J., Moore, S. C., & Brown, G. D. A. (2012). An evolutionary based social rank explanation of why low income predicts mental distress: A 17 year cohort study of 30,000 people. *Journal of Affective Disorders, 136*(3), 882-888. doi: 10.1016/j.jad.2011.09.014

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*(1), 175-180. doi: 10.1037/a0021553

Wood, A. M., Brown, G. D. A., & Maltby, J. (2012). Social Norm Influences on Evaluations of the Risks Associated with Alcohol Consumption: Applying the Rank-Based Decision by Sampling Model to Health Judgments. *Alcohol and Alcoholism, 47*(1), 57-62. doi: 10.1093/alcalc/agr146

Wood, A. M., Brown, G. D. A., Maltby, J., & Watkinson, P. (2012). How Are Personality Judgments Made? A Cognitive Model of Reference Group Effects, Personality Scale Responses, and Behavioral Reactions. *Journal of Personality, 80*(5), 1275-1311. doi: 10.1111/j.1467-6494.2012.00763.x

Yechiam, E., Barron, G., & Erev, I. (2005). The role of personal experience in contributing to different patterns of response to rare terrorist attacks. *Journal of Conflict Resolution, 49*(3), 430-439.

Zhu, S.-H., & Anderson, N. H. (1991). Self-estimation of weight parameter in multiattribute analysis. *Organizational Behavior and Human Decision Processes, 48*(1), 36-54. doi: http://dx.doi.org/10.1016/0749-5978(91)90004-D