ADVISORS AND GROUPS
ESSAYS IN SOCIAL DECISION MAKING

Johannes Müller-Trede

TESI DOCTORAL UPF / ANY 2012

DIRECTOR DE LA TESI

Robin Hogarth, Departament d'Economia i Empresa

UNIVERSITAT POMPEU FABRA

Acknowledgements

Robin Hogarth has been a great teacher during my doctoral studies. A PhD is arguably as much about learning and being trained as a researcher as it is about writing a thesis, and I have learned many things from Robin. Few people inspire as much respect and admiration in me as he does, both as a scientist and as a person. I am deeply grateful for his guidance, for his ability and willingness to listen, and for the time I have spent with him.

I have also learned from many other people whom I would like to thank, including Manel Baucells and Fabrizio Germano, Gert Cornelissen and Gaël Le Mens, Joan Monràs, Rosemarie Nagel, Omiros Papaspiliopoulos and Karl Schlag, as well as Hannes Schwandt and Emre Soyer at UPF, and Judith Avrahami, Maja Bar-Hillel, Shoham Choshen-Hillel, Yaakov Kareev and Ilan Yaniv at the Hebrew University in Jerusalem.

Finally, I would like to thank Laura, Marieke, my friends, and my family, who have not only taught me many things but who have also had to put up with me in different moments during the last five years.

Abstract

The three chapters of this thesis investigate social aspects of judgment and decision making. Chapter One analyses the consequences of making decisions based on predictions of future well-being, and the conditions under which advice can improve these decisions. It shows that an interaction between errors in affective forecasts and the choice process leads to suboptimal decisions and disappointment, and establishes conditions under which advice reduces these effects. The second chapter investigates the boundaries of the result that eliciting more than one estimate from the same person and averaging these can lead to accuracy gains in judgment tasks. It reveals that the technique works only for specific kinds of questions, and people are reluctant to average their initial answers when asked for a final estimate. Finally, Chapter Three reviews experimental results regarding individual and small group behaviour in strategic decision tasks and provides a theoretical framework to analyse the observed differences.

Resum

Aquesta tesi investiga diferents aspectes socials de la presa de decisions. El primer capítol analitza les decisions preses en base a les prediccions del benestar futur, i en quines situacions els consells d'altres persones poden millorar aquestes decisions. Es mostra que una interacció entre el procés de l'elecció i les imperfeccions de les prediccions condueix a decisions subòptimes i a la decepció, i s'estableixen les condicions sota les quals els consells redueixen aquests efectes. El segon capítol investiga els casos en què les persones poden millorar les seves prediccions numèriques donant més d'una estimació i prenent-ne la mitjana. A base d'un experiment, es mostra que la tècnica funciona només amb determinats tipus de preguntes, i que les persones són averses a prendre mitjanes de les seves estimacions inicials quan es pregunta per una estimació final. L'últim capítol revisa els resultats experimentals referents a la presa de decisions estratègiques de la persona individual comparats amb els de la persona que forma part d'un grup reduït i proporciona un marc teòric en el que analitza les diferències que s'observen en el seu comportament.

Preface

The three chapters of this thesis represent the three relatively independent research projects that I have worked on as a PhD student. They investigate different aspects of human judgment and decision making: when other people's advice can help us make better decisions, whether we can make more accurate judgments by thinking of more than one answer to the same question, and how making decisions as part of a group differs from making them individually.

Although these questions may at first seem unrelated, there are two underlying themes which motivate them, and my interests as a behavioural scientist. The first is a concern with social decision making and the aggregation of opinions and preferences. The thesis' title, *Advisors and groups: essays in social decision making*, alludes to this concern, and it manifests itself in each of the three chapters. Advice and advisors protagonise the first chapter, and group decisions the last, but even the second chapter, which is concerned with individual judgments, returns to the issue of aggregation when analysing how people give definite answers to questions that they have answered more than once. The second theme that re-appears throughout the thesis is that decision makers are imperfect but generally unbiased in their judgments. Chapters 1 and 3 show how such imperfect decision makers can be modelled mathematically and how the behaviour observed in experiments relates to these models. Advice and social decision making are also especially relevant when decision makers are imperfect since they may help them make better decisions. Perfect decision makers who always know what to choose do not need anybody's advice!

The order of the chapters reflects the order in which I began to work on each of the projects. The first chapter, *Choice and advice on the basis of affective forecasts*, is based on my Tesina *Random Errors and Systematic Effects*, which I first presented at UPF in 2008. It has accompanied me throughout my doctoral studies- I finished the present version of this chapter in January 2012, and I consider it to be the central chapter of the thesis. It analyses the consequences of making decisions based on predictions of future satisfaction or well-being, *affective forecasts*. I show that an interaction between errors in affective forecasts and the choice process can lead to suboptimal decisions and disappointment and then model how information about other people's affective forecasts and affective reactions -advice- can help people make better decisions. The analysis shows that the usefulness of advice depends on the similarity between tastes of advisors and advisees and the nature and the size of the errors in affective forecasts. Advice is only useful when advisee and advisor share similar tastes, and the best advisors are independent and experienced.

The second chapter, *Repeated judgment sampling: Boundaries* investigates the boundaries of the recent result that eliciting more than one estimate from the same person and averaging these can lead to accuracy gains in judgment tasks. It first examines its generality, analysing whether the kind of question being asked has an effect on the size of potential gains. Experimental results show that the question type matters. Previous results reporting potential accuracy gains are reproduced for year-estimation questions, and extended to questions about percentage shares. On the other hand, no gains are found for general numerical questions. I then test repeated judgment sampling's practical applicability by asking judges to provide a third and final answer on the basis of their first two estimates. In an experiment, the majority of judges do not consistently average their first two answers. As a result, they do not realise the potential accuracy gains from averaging. This chapter was published in *Judgment and Decision Making* in June 2011.

Finally, *On comparing individual and group behaviour in strategic decisions* constitutes the third and last chapter of this thesis. It reviews experimental results regarding individual and small group behaviour in a variety of strategic decision tasks. It provides a theoretical framework for analysing differences between individual and group behaviour which abstracts from complex psychological processes. In particular, it identifies belief updating, aggregation, and social preferences as three dimensions on which decision making as part of a group differs from making decisions individually, and attempts to formalise these. The framework offers a 'rational' account of the experimental finding that groups coordinate better than individuals. At the same time, it suggests that a better understanding of how the group setting affects social preferences is required in order to explain group behaviour in other strategic contexts.

Contents

List of Figures

List of Tables

Chapter 1

CHOICE AND ADVICE ON THE BASIS OF AFFECTIVE FORECASTS

1.1 Introduction

Predictions of our future happiness, of our future satisfaction and our affective reactions are often difficult. Nonetheless, they inform many of the decisions we make. When choosing which film to see at the cinema, people often choose the film that they expect to be the most enjoyable. When choosing what to study at university, high school graduates make predictions about their future satisfaction with both the subject of study and the possible career paths it may open for them, and choose accordingly. In everyday decisions such as what food to order or what film to watch as well as important one-off decisions such as what to study, decision makers engage in the difficult task of imagining future experiences and their reactions to them, and then choose on the basis of these predictions. This paper investigates the consequences of making decisions on the basis of such affective forecasts, and how information about others' affective forecasts and affective reactions can help when making such decisions.

Building on an experimental literature that documents the fallibility of affective forecasts (reviewed in Wilson and Gilbert, 2005), I first integrate recent theories of choice based on anticipated emotions (Kahneman et al., 1997; Mellers et al., 1999) with statistical models of choice in the presence of forecast errors (Harrison and March, 1984; Van den Steen, 2004; Smith and Winkler, 2006). My analysis shows that making decisions on the basis of imperfect affective forecasts can lead to suboptimal choice and post-decision disappointment. Distinguishing between unbiased errors in affective forecasts on the one hand, and psychological biases on the other, Section 1.2 analyses the mistakes we make in affective forecasting and their consequences for the outcomes of our choices.

I then investigate how advice in the form of information about other people's affective forecasts and affective reactions may help make better choices, modeling advice as additional affective predictions. Just as the estimates of others can help people make more accurate judgments (Yaniv, 2004; Surowiecki, 2004; Soll and Larrick, 2009), better decisions can be made by taking into account others' affective forecasts and affective reactions. My model also suggests how advice-taking in choice may differ from advice-taking in judgment. It predicts that the usefulness of advice depends on whether advisors' tastes resemble those of their advisees, as well as the nature and the size of the forecast errors. The analysis in Section 1.3 suggests that advice can help if advisor and advisee are sufficiently similar to one another, but the prediction errors they make in their affective predictions are not.

1.2 Imperfect affective forecasts

This section analyses the different mistakes decision makers incur when making affective forecasts, and how they interact with the choice process. I distinguish between random- and systematic errors in affective forecasts, and discuss differences between them as well as their similarities, both in terms of their effects and in terms of the environments which give rise to either type of error. Throughout this section, I am concerned with a single decision maker. The results of this analysis will then serve as a basis for discussing the usefulness of advice in Section 1.3.

1.2.1 Choices based on affective forecasts

The idea that we choose on the basis of predictions of our emotional reactions to the outcomes of our actions has a long history (see Mellers, 2000, for a review). Recently, it has received renewned attention in the form of *Subjective Expected Pleasure Theory* (Mellers et al., 1997, 1999) on the one hand, and the distinction between *decision utility* and *experienced utility* explored by Kahneman and his co-authors (Kahneman et al., 1997; Kahneman and Thaler, 2006) on the other. Both theories distinguish between the affective reaction to an outcome once it is realised, and the decision maker's prediction of this affective reaction before it is realised, the affective forecast. By making this distinction, they allow decision makers to suffer from *miswanting*, that is, to want something more than they actually like it (Gilbert and Wilson, 2000). Subjective

Expected Pleasure Theory explains how the characteristics of the choice situation shape affective forecasts by incorporating outcome, surprise, regret and disappointment effects as well as reference points (Mellers et al., 1999). Kahneman and Thaler (2006) show how psychological biases can lead to imperfect affective forecasts.

The present analysis adds to these considerations by highlighting the interaction between errors in affective forecasts and the choice process, and explicitly analysing its effects. Following Harrison and March (1984) and Smith and Winkler (2006), I model decision makers as choosing between n alternatives which yield levels of pleasure $\mu_1, ..., \mu_n$. This *pleasure* reflects the hedonic experience or affective reaction the decision maker would experience upon choosing a particular alternative and is unknown at the time of choice. Instead, decision makers choose on the basis of the imperfect *affective forecasts* $V_1, ..., V_n$. 'Doing what is best for oneself' can then be thought of as choosing that alternative i^* with the maximal V_{i^*}. I analyse the expected difference $\mathbb{E}[\mu_{i^*} - V_{i^*}]$ between the pleasure associated with the chosen option i^* and its affective forecast across many such decisions. This expected difference, the so-called post-decision surprise (Harrison and March, 1984) associated with the chosen option, can be seen as a measure of *miswanting*.

1.2.2 Unbiased errors

Decision makers who are able to predict their tastes and affective reactions on average but may not be able to predict them perfectly can be modelled as incurring random, unbiased errors in their affective forecasts. Such errors imply that in expectations, pleasures μ and affective forecasts V are equal for any given alternative so that $\mathbb{E}[V_i|\mu_1, ..., \mu_n] =$

$\mu_i \; \forall i$. Why should our affective forecasts suffer from such random errors? At least two important sources of uncertainty in decision making can be modelled by means of such unbiased errors: uncertainty regarding the properties of the alternatives, including uncertainties regarding the possible outcomes that may exist[1], and uncertainty regarding the decision maker's preferences (cf. Loomes et al., 2009, 's distinction between extrinsic and intrinsic uncertainties). Many decisions we make are subject to such uncertainties, and unless there are good reasons to expect them to create biases in our perceptions (and Section 1.2.4 shows there sometimes are), they can be thought of as inducing unbiased randomness in our decisions.

Choosing which movie to watch on a streaming website or at the cinema is a good example in which both of these uncertainties are likely to be present. Since novelty is part of the fun when it comes to movies, we watch most of them only once, and a choice between films is usually a choice between unknown alternatives. When choosing, we are necessarily uncertain about the quality of the acting or the details of the scriptwriting. In addition, even if we had detailed information about the film, we may still not always be able to predict our affective reaction correctly. Affective reactions to pieces of art (such as films) are hard to predict because of the complexity of the object, and are influenced by our state of mind. If you are not able to anticipate that you are not in

[1]In complex decisions such as deciding between jobs or university degrees, for example, a decision maker may not be able to enumerate all the different possible outcomes associated with taking a particular action. Such decisions are difficult to analyse with models of expected utility and even models of ambiguity, in which decision makers may have imperfect or no information about the probabilities associated with the different outcomes, but still know *which* outcomes are possible (Einhorn and Hogarth, 1986; Gilboa and Schmeidler, 1989).

the mood for a comedy, an evening at the cinema can be disappointing even though the same film would have made you laugh on another day.

Another important class of decisions in which both the properties of the alternatives and our own preferences are likely to be uncertain is that of intertemporal decisions. Even simple choices like for example the choice between a chocolate bar and an apple can become complex when an (inter)temporal dimension is added. While the correct choice may be clear when deciding for immediate consumption, knowing whether chocolate would be preferable to an apple two weeks from now requires knowing what the weather will be like and what one will eat for lunch or breakfast two weeks from now, as well as one's future state of mind. Generally, random mistakes in affective forecasts may provide a means to modeling preferences which are not yet well-established.

1.2.3 The Optimizer's Curse

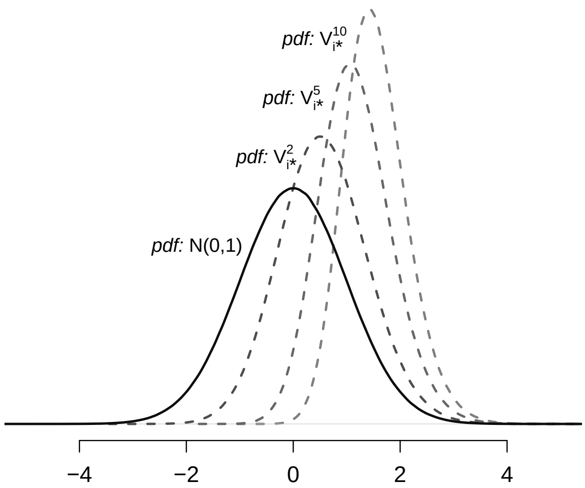
What are the implications of unbiased errors for the degree to which people *miswant*? One reason why unbiased errors in affective forecasting have not received much attention in the literature could be that they seem innocuous. If decision makers' forecasts are unbiased, why should the errors have any behavioural effect? Harrison and March (1984), however, show that even unbiased errors can lead to disappointment. Although the post-decision surprise associated with any particular option i is zero, so that $\mathbb{E}[\mu_i - V_i] = 0$ for all i, the post-decision surprise associated with the *chosen* option is negative: $\mathbb{E}[\mu_{i^*} - V_{i^*}] \leq 0$. Smith and Winkler (2006) refer to this phenomenon as the *Optimizer's Curse*, because the expected disappointment does not arise from a property of the errors but from the interaction of the errors with the maximisation process instead.

For an example of the Optimizer's Curse, consider the cinephile Claire, who has decided to stay in and watch a film on a Wednesday night. Her flatmate, Pauline, who is also fond of arthouse cinema, has recently bought n new films, and Claire wants to watch one of them. For the purpose of the illustration, assume that all films are equally enjoyable, so that the pleasure associated with them can be standardised to zero: $\mu_i = 0 \; \forall i$. Claire's affective forecasts are imperfect, and I assume them to be independently and Normally distributed with mean zero and a standard deviation of one: $V_i \sim N(0, 1) \; \forall i$. Her affective forecasts are not biased: on average, Claire will predict the pleasure associated with the films correctly since $\mathbb{E}[V_i] = \mu_i = 0 \; \forall i$. Based on these unbiased affective forecasts, she watches the film i^* that promises to be most enjoyable, so that $V_{i^*} = \max(V_i)$.

Consider first the case where $n = 1$. Of all the new films, there is only one Claire has not yet seen. Her affective forecast of the pleasure that watching the film would bring is distributed as $V_1 \sim N(0, 1)$. What is the affective forecast associated with the chosen option? Since $V_{i^*} = \max(V_1) = V_1$, it follows that $V_{i^*} \sim N(0, 1)$, and hence that $\mathbb{E}[V_{i^*}] = \mu_1 = 0$. When she has no choice, Claire's affective forecasts regarding the pleasure associated with the chosen film are unbiased.

However, this is not the case if Claire has more than one film to choose from. Assume that $n = 2$, and $V_1, V_2 \sim N(0, 1)$. Then $V_{i^*} = \max(V_1, V_2)$, and Claire's choice is determined by the *larger* affective forecast. The smaller of the two, on the other hand, will not have any effect on the choice. This phenomenon can be thought of as a sampling bias caused by the maximisation routine: positive errors in affective forecasts make films appear more enjoyable and therefore more likely to be chosen. Negative errors, on the other hand, make a film look less appealing, and Claire will be less likely to watch it. Mathematically, this sampling bias implies that for $n \geq 2$, the distribution of V_{i^*} is not standard Normal anymore. Figure 1.1, adapted from Smith and Winkler (2006), compares the distribution of the V_i with the distribution of V_{i^*} for different values of n.

Figure 1.1: The distribution of the maximum of n standard Normal affective forecasts, $V_{i^*}^n$



Maximising expected subjective pleasure then implies that a choice between n options is determined by the largest of the n affective forecasts associated with them. Figure 1.1 shows that for $n \geq 2$, this sampling bias shifts the distribution of V_{i^*} to the right compared with the standard Normal distribution that the V_i are drawn from[2]. As a result, Claire overestimates the pleasure associated with her chosen film on average: $\mathbb{E}[V_{i^*}] > 0 = \mu_{i^*}$. She then perceives her chosen film as disappointing on average since the expected post-decision surprise associated with the chosen option is negative, $\mathbb{E}[\mu_{i^*} - V_{i^*}] < 0$. This effect, which arises from the interaction between forecast errors and the maximisation routine, is what Smith and Winkler (2006) call the Optimizer's Curse. It is more pronounced for larger choice sets, since adding an alternative can only lead to greater disappointment: if the new alternative appears more enjoyable than the other available alternative, disappointment increases, and if it does not, disappointment remains unchanged. Even for $n = 2$, the expected disappointment is as large as 56% of a standard deviation of the errors, however[3].

This example involving equally good alternatives illustrates the Optimizer's Curse nicely but is peculiar in that the decision maker is indifferent as to which option to choose. Errors in affective forecasting become more interesting in choice situations where there are "right" and "wrong" decisions. Choosing the right film could mean ninety minutes of brilliant entertainment, and choosing the wrong one falling asleep on the sofa.

What are the effects of errors in affective forecasts in a choice between options which yield different levels of pleasure? To answer this

[2]The analyses in this paper are based on the results of Monte Carlo simulations programmed in the R environment for statistical computing. The simulations are available from the author on request.

[3]For $n = 2$, $\mathbb{E}[\mu_{i^*} - V_{i^*}] = -.56$. As Smith and Winkler (2006) point out, this finding is invariant to scale and location, and can be generalised to other Normal distributions, so that the expected disappointment in a choice between two equally good options with Normal forecast errors is 56% of a standard deviation of these errors.

question, I model pleasure levels as drawn from a Normal distribution. This is for illustration purposes only: pleasure levels are determined by the underlying theory such as subjective expected pleasure theory, and any one realisation of μ_1 and μ_2 can be thought of as resulting from such a theory. Modeling pleasure levels stochastically has an important advantage for the exposition, however, because the resulting framework can readily incoporate systematic biases as well as second opinions and will serve as a basis of comparison for the remainder of this chapter.[4]

Assume, then, that in a choice between two options, pleasure levels are determined as $\mu_1, \mu_2 \sim N(0, \sigma_\mu^2)$. As before, affective forecasts are imperfect, and are drawn from Normal distributions centered on the pleasure levels μ_1 and μ_2, so that $V_1 \sim N(\mu_1, \sigma_V^2)$ and $V_2 \sim N(\mu_2, \sigma_V^2)$. The mechanics of the choice process also remain unchanged, and the option with the highest affective forecast is chosen: $V_{i^*} = \max(V_1, V_2)$. All that has changed with respect to the first example is that pleasure levels are now not fixed anymore and can differ between the alternatives. The degree to which they can be expected to differ is given by the variance σ_μ^2, and similarly, the expected impact of the errors in the affective forecasts is captured by their variance, σ_V^2. Choice environments can then be characterised by the variance ratio $\frac{\sigma_V^2}{\sigma_\mu^2}$. This ratio is large if either errors are large or pleasure levels vary little between options, such as in Claire's choice between promising arthouse films. When preferences are well-determined or prediction errors are small, the variance ratio will

[4]Second opinions and advice are especially hard to incorporate in classic theories of choice: while similarity between random variables is measured easily (by correlations, for example), it is not obvious how to measure similarities between (experienced) utility functions, although common sense would suggest that some people's preferences are more similar than others'.

be small. For an example, think of Claire choosing between this year's winner of the Golden Palm at the Cannes film festival and a Hollywood blockbuster. Figure 1.2 shows how expected disappointment and the probability of choosing suboptimally vary with the variance ratio $\frac{\sigma_V^2}{\sigma_\mu^2}$ in a choice between two options.

Figure 1.2: Expected disappointment in a choice between two unequal options

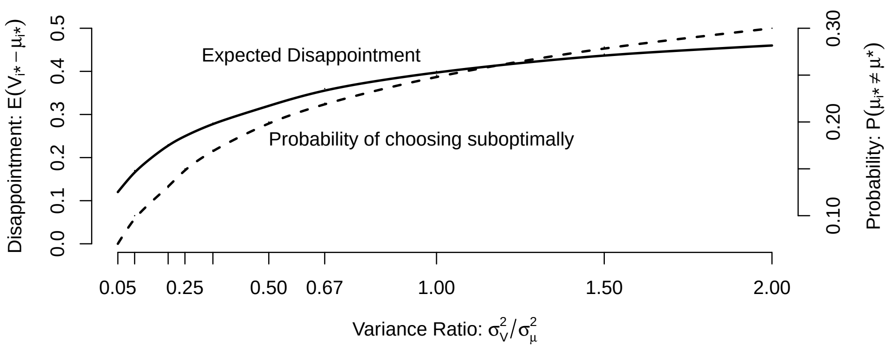


Figure 1.2 confirms the intuition that imperfect affective forecasts should have a larger effect when the errors are large, or pleasure levels do not differ much between alternatives. Expected disappointment rises from 12% of a standard deviation of the forecast errors at a variance ratio of .05 to 46% for a variance ratio of 2. Figure 1.2 also graphs the probability of making a suboptimal choice. Remember that when pleasure levels differ between the two options, a 'better' and a 'worse' option exist. Denoting the maximal pleasure associated with any of the options by μ^*, the right axis of Figure 1.2 charts the probability that the pleasure associated with the chosen option is not the maximal pleasure the decision maker could have obtained, so that $\mu_{i^*} \neq \mu^*$. The dashed

line shows that, like expected disappointment, the probability that the worse alternative is chosen rises as errors get larger or pleasure levels more similar.

1.2.4 Systematic errors

When people consistently make the same mistake in forecasting their affective states, the error can be thought of as a bias. While decision makers are on average correct in their affective forecasts when their errors are unbiased, they are on average incorrect if these are biased. Biases in predicting future affective states have received considerable attention in the literature, and have been reviewed in Gilbert and Wilson (2000) and Wilson and Gilbert (2005). I will briefly discuss the psychological biases known as impact-, projection- and diversification bias as well as attitudinal biases and then model them in the mathematical framework introduced above.

Impact bias makes people overestimate the intensity and duration of their affective responses to future events. Examples include assistant professors who overestimate the emotional impact that a positive or a negative tenure decision will have on their lives and voters who mistakenly believe that the outcome of a political election will have a lasting impact on their happiness (Gilbert et al., 1998). Even with respect to life-changing experiences, we exhibit impact bias, like the women in Mellers and McGraw (2001)'s study who expect to feel worse about the unwanted outcome of a pregnancy test than they actually do. Impact bias causes people to exaggerate in their affective forecasts, both the exhilaration of a positive experience and the despair of a negative one.

Note that while random errors arise from uncertainties regarding the alternatives and the decision maker's preferences over them, impact bias is a result of *how* we decide. Take the participants in Schkade and Kahneman (1998)'s study who think that people who live in California are happier than those who live in the Midwest, for example. When imagining what life would be like in California, Midwesterners may think of beautiful Pacific beaches and sunny afternoons. In their minds, they focus on the geographical difference and neglect other factors such as interpersonal relations or professional success and their effect on life satisfaction. Such misguided *focalism* (Schkade and Kahneman, 1998; Wilson et al., 2000) leads to impact bias as it makes us overestimate the relative importance of a particular feature of an outcome and its effect on our happiness.

A second cause for impact bias is what Gilbert et al. (1998) refer to as *immune neglect*. People underestimate the degree to which our "psychological immune system" can protect us from lasting effects of negative experiences on our well-being. Negative experiences, including the denial of tenure and the unwanted outcomes of an election or a pregnancy test, hurt, but only for a while. Immune neglect refers to our inability to predict correctly how long this disappointment will last.

Impact bias is not the only systematic mistake we make when predicting future affective states. At least two other psychological biases exist. *Projection bias* refers to our tendency to anchor too much on our current emotional, physical and motivational states when forecasting the affective states we might be in in the future (Loewenstein et al., 2003; Read and Van Leeuwen, 1998). *Diversification bias* describes how we seek excessive variety when choosing simultaneously for sequential consumption (Read and Loewenstein, 1995; Simonson, 1990). Although I will refrain from discussing the underlying processes in greater detail (see Kahneman and Thaler, 2006, for an overview), like impact bias they are systematic and arise from the interaction of the choice environment and psychological processes.

Finally, biases in affective forecasts can also arise from stereotypes and attitudes. Attitudes can be hard to change (Wood, 2000; Sherman and Cohen, 2002), and even more so in social contexts and in the presence of social influence (Wood, 2000; Crano and Prislin, 2006). When shared attitudes lead to shared mistakes in affective forecasts, these can be interpreted as biases. While not as general as psychological biases, which are general characteristics of human decision making, they can have the same effect if people interact mostly with others who are similar to them. In some cases, the attitudinal bias may even be shared by virtually all of society. New technologies which affect people's social interaction come to mind: Internet matchmaking sites, for example, made many people feel uncomfortable when they were first introduced. I will therefore treat attitudinal biases as systematic biases in people's affective forecasts.

1.2.5 Systematic errors and post-decision surprise

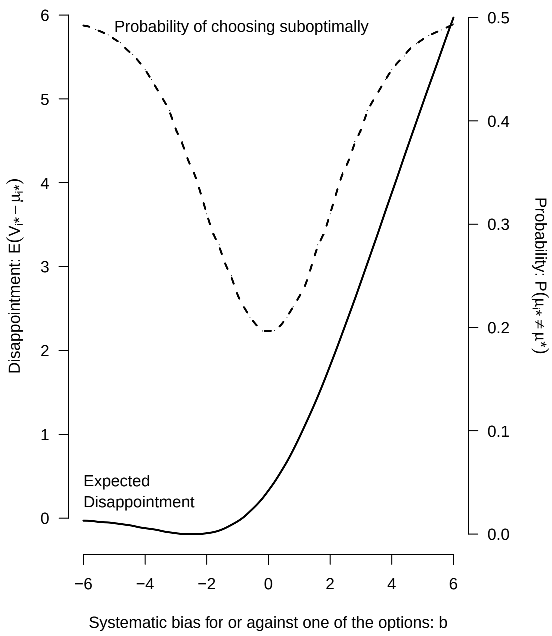
In terms of values μ and affective forecasts V, systematic biases can be modelled as $V_i = \mu_i + b_i$, where b_i represents the bias. Positive values of b_i indicate that decision makers systematically overestimate the satisfaction associated with option i, and negative ones indicate that they underestimate it. Similarly, random- and systematic errors can be combined such that $\mathbb{E}[V_i|\mu_1, ..., \mu_n] = \mu_i + b_i$ since systematic errors in affective forecasts will often be accompanied by additional uncertainties causing random errors.

What is the effect of systematic errors in affective forecasts on post-decision surprise? Consider again the setting of a choice between two options. As before, pleasure levels are drawn from a Normal distribution, $\mu_1, \mu_2 \sim N(0, \sigma_\mu)$ and random errors are still distributed Normally around them, but in addition, the decision maker is now systematically biased for or against the first option so that $b_1 = b$ and $b_2 = 0$. The affective forecasts for the two options are then $V_1 \sim N(\mu_1 + b, \sigma_V)$ and $V_2 \sim N(\mu_2, \sigma_V)$ and the option with the higher affective forecast is chosen so that $V_{i^*} = \max(V_1, V_2)$. How the bias affects expected disappointment and the probability of choosing suboptimally is shown in Figure 1.3.

Two assertions can be made on the basis of Figure 1.3. First, the symmetry of the dashed line with respect to the case where $b = 0$ shows that in a choice between two options, the probability of choosing the worse of the two is determined by the absolute size of the bias, not its sign. Second, the relationship between expected disappointment and the bias is more complicated and asymmetric with respect to $b = 0$.

I begin by analysing the probability of choosing suboptimally. Section 1.2.3 shows that in the absence of a systematic bias ($b = 0$), this probability is determined by the variance ratio $\frac{\sigma_V^2}{\sigma_\mu^2}$ which summarises the importance of the errors in a choice setting. In Figure 1.3 the variance ratio is .5, so errors are large: even without a bias, the decision maker chooses the worse option in about 20% of cases (cf. Figure 1.2). The graph shows that biases further make things worse, since the dashed

Figure 1.3: Expected disappointment in a choice between two options with a bias ($\frac{\sigma_V^2}{\sigma_\mu^2} = .5$)



line increases both with a negative and with a positive bias. At $b \pm 6$[5], the probability of choosing suboptimally rises to almost 50%. Why fifty per cent? Since $\mu_1, \mu_2 \sim N(0, \sigma_\mu)$, the probability that either option is better (or worse) is $\mathbb{P}[\mu_1 > \mu_2] = \mathbb{P}[\mu_2 > \mu_1] = .5$. Now if a positive bias is large enough, it completely determines the choice and the biased op-

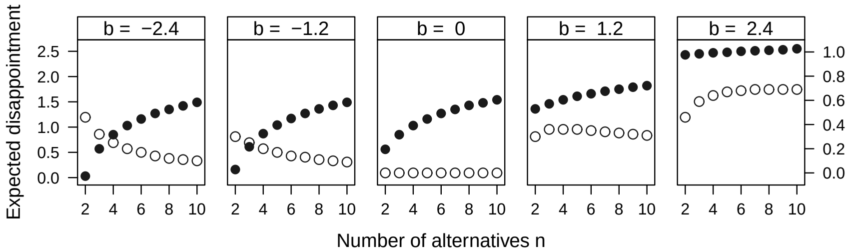
[5]The variance ratio $\frac{\sigma_V^2}{\sigma_\mu^2}$ also determines what constitutes a 'large' value, of course. Here, 'large' means that the bias is large enough to completely determine the outcome of the choice, that is, that the probability of random errors overcoming the effect of the bias is zero. For Normally distributed μ and errors and a variance ratio of .5, a value of 6 is large as can be computed from the properties of the Normal distribution.

tion will always be chosen. Since the probability that the biased option is better than the alternative is .5, in the other 50% of cases, choosing the biased option will be suboptimal. Similarly, if a negative bias is large enough, the biased option is never chosen, and an analogous argument holds.

Consider now the implications for the expected disappointment associated with the choice. Note that the expected disappointment associated with the biased option is $\mathbb{E}[V_1 - \mu_1] = \mathbb{E}[\mu_1 + b - \mu_1] = b$, and that of the unbiased option is $\mathbb{E}[V_2 - \mu_2] = 0$. If a large positive bias implies that the biased option is almost always chosen, the expected disappointment associated with the chosen option will be equal to that of the biased option so that $\mathbb{E}[V_{i^*} - \mu_{i^*}] = \mathbb{E}[V_1 - \mu_1] = b$. This explains the near linear relationship between the size of the bias and expected disappointment for $b > 0$. On the other hand, the effect is different when the bias is negative. A large negative bias implies that the unbiased alternative is always chosen, so that $\mathbb{E}[V_{i^*} - \mu_{i^*}] = \mathbb{E}[V_2 - \mu_2] = 0$. Only medium-sized negative biases which still allow for the biased option to sometimes appear superior (so that decision makers choose this option and subsequently realise that they underestimated the pleasure associated with it), lead to *negative* expected disappointment. Generally, while decision makers are likely to act upon positive biases and consequently be disappointed, they are unlikely to choose a course of action they are negatively biased against, and are hence unlikely to be positively surprised. The asymmetry induced by the maximisation process which gives rise to the Optimizer's Curse when errors in affective forecasts are unbiased therefore also affects the outcome of choices in the presence of positive or negative biases.

What does this imply for choices between more than two alternatives? For an illustration, let us return to Claire's choice between n equally pleasurable films with $\mu_i = 0 \; i = 1, ..., n$. Claire is at a video rental store and wants to choose two films which she plans to watch later that day. She has narrowed down her choice set to "The complete works of Stanley Kubrick" and one other film by a different director. She decides that she wants to watch Kubrick's *Dr. Strangelove*, and diversification bias now makes Claire overestimate the pleasure associated with watching this other film relative to the remaining Kubrick films. Her affective forecasts can then be modelled as $V_1 \sim N(b, 1)$ and $V_i \sim N(0, 1) \, \forall i \neq 1$. For different degrees of the bias b, Figure 1.4 shows how Claire's expected disappointment varies with the number of films n[6].

Figure 1.4: A choice between n equally good alternatives in the presence of biases of different sizes



Expected disappointment is not symmetric with respect to the unbiased case when $n = 2$, and the black dots reveal that this asymmetry between between positive and negative biases also exists for larger choice

[6]Negative values of b could be illustrated analogously by a situation in which Pauline had bought n films of which two were by the same director, and having watched one of them, Claire is now biased against the other film by the same director when compared to the rest of the films because of diversification bias.

sets. A sufficiently large positive bias will dominate the outcome of the choice even if n is large. The rightmost panel of Figure 1.4 shows how little expected disappointment varies with n in the presence of a large positive bias, $b = 2.4$[7]. This echoes the previous finding that expected disappointment can be determined almost exclusively by the bias. The first two panels of Figure 1.4 paint a different picture for negative biases, however. The black dots in these panels behave much like those in the middle panel which shows the expected disappointment incurred from the Optimizer's Curse alone ($b = 0$). A difference is only evident for very small choice sets: for $n = 2$ expected disappointment is smaller in the presence of a negative bias, and almost zero for $b = -2.4$. For choice sets with more than three alternatives, however, a negative bias has almost no effect on expected disappointment.

The white circles in Figure 1.4 reveal why this is the case. On the right axis, they represent the probability that the outcome of the choice would be different if there was no bias. While negative biases lose their effect on the outcome of the choice quickly as the number of alternatives increases, the effect of a positive bias increases with n intially, and subsequently diminishes at a slower rate. Like the Optimizer's Curse, this asymmetry arises from the maximisation process. A sufficiently strong positive bias reduces the choice set to a single option: the biased option is almost always chosen. On the other hand, the strongest effect a negative bias can have is to remove one option from the choice set. In the presence of a large negative bias, the biased option is almost never chosen. Positive biases are therefore more likely to have an effect

[7]Since $\mu_i = 0 \; \forall i$ and $\sigma_V = 1$, $b = 2.4$ is now 'large' since draws from a standard Normal distribution only exceed 2.4 in 0.8% of cases.

on the outcome of our choices than negative ones, just like a positive realisation of a random forecast error is more likely to affect our choice than a negative one. In terms of expected disappointment, this means that a choice between n options where one option exhibits a large negative bias is equivalent to choosing between $n-1$ unbiased options. In a choice between n options with a large positive bias, on the other hand, expected disappointment is determined largely by the size of this bias.

1.2.6 Discussion

In this section, I have argued that decision makers who aim to maximise their well-being on the basis of affective forecasts are prone to experience disappointment if these forecasts are imperfect. This potential disappointment is a product of the interaction of the forecast errors with the maximisation process, and can be observed for both random or unbiased errors, and for systematic biases in affective forecasts. Mathematically, maximisation processes are more sensitive to positive errors in the estimates of the quantity that is being maximised than to negative ones. Intuitively, errors can make the options in a choice situation look more appealing or less appealing than they really are- and if you choose things or actions, or films, or yoghurts, or opera performances because they look appealing, you are more likely to choose that which looks more appealing than it is than the alternative which looks less appealing than it is.

These findings imply that since both systematic and random errors can cause disappointment and decision makers may not be aware of this, observing systematic post-decision disappointment does not allow us to make an inference about the nature of the errors. This has methodological implications for empirical and experimental research. As an example, consider the experimental research on diversification bias. In the experimental setup introduced by Simonson (1990), undergraduate students had to make three choices from a menu of six different snacks, to be consumed after class over a period of three weeks. Diversification bias refers to the finding that those students who had to make all three choices simultaneously at the beginning of the first week chose more different snacks and reported lower choice satisfaction than students who could choose sequentially at the time of consumption. Such choice patterns -more different choices, and lower satisfaction- could be caused by a bias (see Read and Loewenstein, 1995), but are also consistent with an interpretation in which the affective forecasts involved in simultaneous choice are subject to random errors. The disappointment may then partly be caused by random errors, an idea which is closely related to one of the early interpretations of the variety-seeking behaviour observed in these studies (see Kahn, 1995)[8]. Such an account of diversification bias is reminiscent of "sampling" explanations of other biases including illusory correlations (Denrell and Le Mens, 2011), in-group bias (Denrell, 2005; Fiedler, 2000), overconfidence (Juslin et al., 2007; Einhorn and Hogarth, 1978), social influence (Denrell and Le Mens, 2007) and risk aversion (Denrell, 2007; Hertwig et al., 2004; March, 1996) in at-

[8]Read and Loewenstein (1995) include a control condition in which participants can sample the snacks before deciding. While this control condition alleviates concerns about uncertainties regarding the actual snacks, it cannot control for uncertainties regarding one's preferences in a week's time, or in two week's time. Read and Loewenstein (1995) and Read et al. (2001) provide evidence for the psychological processes of time contraction and choice bracketing, and I do not wish to argue that random errors are the main explanation for diversification bias. The strength of the systematic effect is hard to assess in a design in which maximisation with unbiased errors in affective forecasts predicts an effect in the same direction, however.

tributing the observed bias to the properties of the choice environment and the choice process rather than a cognitive malfunction.

As a model of disappointment, this research is also related to *disappointment theory* (Bell, 1982; Delquie and Cillo, 2006; Loomes and Sugden, 1986) on the one hand, and *regret theory* (Loomes and Sugden, 1982) on the other. The disappointment here arises from the difference between the expectations a decision maker has before taking an action, or affective forecasts, and the affective reaction once the outcome of the same action is realised, however. Its sources are therefore different from those in disappointment theory, in which decision makers are disappointed because they compare realised outcomes with unrealised outcomes, or regret theory, in which decision makers are disappointed because they compare the outcome of chosen actions with those of foregone actions.

A closer connection exists between the present research and the family of models known as random utility models (Block and Marschak, 1960; Becker et al., 1963; Loomes, 2005). These models are mathematically very similar to the present model but do not differentiate between affective forecasts and realised pleasure, and view the stochastic component of choice as either determined by preferences which can vary with time (Thurstone, 1945) or as uncertainty arising from the model, not the choice itself (Manski, 1977). As a result, random utility models have prviously not been related to post-decision disappointment. Future experimental research should examine how the different notions of disappointment and regret interact, as decision makers may simultaneously experience the disappointment from realising that their chosen option is not as pleasurable as they expected, and regret from not having chosen a different option, for example.

Finally, I am abstracting from two important questions in this analysis, which also warrant future research. The first of them concerns the degree to which decision makers are aware of their errors. The answer depends, I believe, on the choice situation and on what causes the errors. Unbiased errors are largely caused by uncertainties, and it seems plausible that decision makers have a notion of the degree of uncertainty involved in a choice situation. When choosing which film to watch, decision makers know that their knowledge of the films is not perfect. Similary, somebody who has to accept or reject a job offer is aware of not being able to make perfect predictions of the affective reactions to everything the new job entails. It is easily conceivable that decision makers even adapt their behaviour if they are aware of imperfections in their affective forecasts. In choices based on multiple attributes, for example, Hsee (1996, 2000) finds that decision makers put relatively less weight on those attributes which are harder to evaluate. The present framework could be a useful tool for modeling behaviour according to this *Evaluability Hypothesis*, with relative weights on the attributes depending on their predictability as captured by σ_V^2.

Sometimes, an awareness of the imperfections in our own affective forecasts may even make us aware of the Optimizer's Curse they cast over us. A friend of the author, for example, loves to go to the cinema but refuses to pick the film himself to save himself from disappointment. Popular wisdom which tells us to control our expectations may reflect wise people's intuitions about the Optimizer's Curse. Alternatively, people may be aware of the random errors they commit in their affective forecasts without understanding their consequences. This is consistent with people's inability to detect regression effects (Tversky and Kahneman, 1974) and Fiedler and Juslin (2006)'s notion of the decision maker as a naïve statistician. As for systematic errors in affective forecasts, it seems less likely that decision makers could be aware of them. Although creating an awareness of psychological biases such as impact bias has been shown to have little, if any, effect on the accuracy of affective forecasts (Riis et al., 2005; Ubel et al., 2005), there is no evidence that people were aware of the biases beforehand, either (see also Gilbert et al., 2009).

Second, I have not adressed the role of experience explicitly. This omission is intentional and stems from what I consider to be the focus of my analysis: choices based on affective forecasts in environments in which these forecasts are imperfect. In a Bayesian extension to the framework proposed here, realised pleasure levels could be thought of as posteriors on the basis of which decision makers construct and update priors for affective forecasts when decisions are made repeatedly. It should be noted that accommodating for the context-dependence and the fickle nature of hedonic experiences in such a model could be difficult, however. In the remainder of this chapter, I refrain from doing so, but consider experience once more when I investigate how other people's experience affects their usefulness as advisors.

1.3 Advice

Thus far, I have shown that when making choices on the basis of affective forecasts, decision makers may choose suboptimal options and experience post-decision disappointment. In what follows, I analyse how advice -information about other people's affective forecasts and affective reactions- can help decision makers reduce these negative consequences of their choices. I examine the parallels between a decision maker predicting a future affective reaction and a forecaster making a numerical estimate. Drawing on the literature on combining estimates and numerical judgments, I then explore how to combine one's own affective forecast with a piece of advice to make better decisions. The resulting framework for advice-taking in decision making relates advice and its value to the "Wisdom of the Crowd" in forecasting and estimation problems. It shows how the averaging principle implies that decision makers can benefit from taking into account others' advice when their affective predictions are imperfect, but also reveals that advice will only be useful if advisee and advisor are similar to one another. Finally, I relate these results to existing experimental work on advice-taking in decision making.

1.3.1 What is advice?

Advice can be of many forms. Advisors can be experts who share their knowledge with an advisee, for example. Other advisors tell advisees explicitly what to do. This analysis is concerned with a third form of advice: information about other people's affective forecasts or affective experiences. Claire, for example, may know from previous discussions how much Pauline expects to like a certain film. Or, if Pauline has already seen a particular film, she may have told Claire how much she enjoyed watching it. Claire then has access to an additional *affective valuation*: her advisor's affective forecast or a report of the advisor's affective experience. This is potentially valuable information for her which she can incorporate into her affective forecasts regarding her own affective experience of the films.

Friends and colleagues such as Claire and Pauline often share their expectations and their affective reactions. I focus on advice as other people's affective forecasts and experiences, however, because such information is much more readily available. Internet-based social networking technologies, for example, can be seen as platforms for sharing opinions and valuations. Internet retailing businesses employ on-line rating systems, encourage their customers to report how satisfied they are with their purchases, and then make personalised recommendations on the basis of such ratings and customers' purchase histories. These satisfaction ratings aim to capture an affective experience. Information about other people's affective valuations is ubiquitous, and consumers can make better purchases, businesses better recommendations, and decision makers generally can make better decisions when they know how to use this information.

How can advice-taking on the basis of such information be modelled? Let us return to Claire's choice between movies where Claire is the advisee and Pauline the advisor. Claire's preferences can be described by the basic framework introduced above in which pleasure levels are distributed Normally such that $\mu_i^1, \sim N(0, \sigma_\mu)$, with the superscript denoting the individual (Claire) and the subscript the film. Similarly, Pauline's preferences can then be modelled as $\mu_i^2, \sim N(0, \sigma_\mu)$. As before, Claire does not know the pleasure levels μ_i^1 associated with the films when making the decision, and instead has to rely on her affective forecasts V_i^1. In addition, she now has access to Pauline's affective valuation, V_i^2. Having talked to her friend, Claire can decide which film to watch on the basis of both V_i^1 and V_i^2. How should she make this decision?

1.3.2 Combining affective forecasts

Previous models of choice based on affective forecasts (Kahneman et al., 1997; Mellers et al., 1999) have not asked the question as to how affective valuations can be combined because they consider a single decision maker only. A related question which has received a lot of attention, on the other hand, is how to combine different predictions or forecasts when estimating an unknown, objective quantity such as the amount of calories in a food item or the year in which a particular historical event took place (Armstrong, 2001; Galton, 1907; Soll and Larrick, 2009; Stewart, 2001; Surowiecki, 2004; Yaniv, 2004). In such a setting, averaging different judges' estimates usually leads to an improvement in judgment accuracy when compared to an individual judge, a phenomenon often referred to as the "Wisdom of the Crowd". Can crowds also be wise when predicting affective experiences?

To answer this question, compare Claire's choice between films and a judge predicting an unknown quantity. Claire has to make her decision on the basis of two affective valuations, her own and that of her advisor Pauline. The girls form a small crowd of just two people, comparable to participants in the experiments of Yaniv and Kleinberger (2000) or Larrick and Soll (2006) who investigate how people revise their factual and numerical estimates on the basis of those of another person. In such choices, decision makers can be thought of as revising their initial affective forecast on the basis of the advisor's affective valuation, so that their final affective forecast of their enjoyment of a film i is given by

Equation 1.1.

$$V_i^{ad} = \psi * V_i^1 + (1 - \psi) * V_i^2 \tag{1.1}$$

Equation 1.1 implies that the decision maker takes a weighted average of the two affective valuations, a model commonly used in advice-taking in judgment, and in opinion revision. Its interpretation is similar to analogous equations in judgment contexts: the equation models how decision makers translate others' affective forecasts such as "I think the film will be very enjoyable" into their own scales of expected pleasure, and how much weight they attach to such second opinions. Asking whether the "Wisdom of the Crowd" can help decision makers predict their affective experiences is then equivalent to analysing how much weight, if any, a decision maker should place on the advisor's affective valuation in Equation 1.1[9].

What gives rise to the "Wisdom of the Crowd" in estimation tasks is that prediction errors between judges are usually at least partly independent (see e.g. Herzog and Hertwig, 2009; Soll and Larrick, 2009). How does this relate to a choice between films? Claire's and Pauline's affective forecasts are imperfect, just like the judgments of the participants in the experiments of Yaniv and Kleinberger (2000) and Larrick and Soll (2006). Although their prediction errors may not be perfectly independent- reasons are discussed below-, they will often be partly independent. This suggests that gains from averaging different affective valuations could potentially exist. Unlike the judges in an estimation

[9]Equation 1.1 is intuitively appealing and widely used in the literature in combining opinions, but is not necessarily the optimal method of combining affective valuations. An interesting subject for future research would be a formal analysis of the problem in the spirit of Winkler (1981) and Morris (1974, 1977).

task, Claire and Pauline are estimating different quantities, however. Claire predicts μ_i^1, and Pauline predicts μ_i^2: while the amount of calories in an apple does not change with the person estimating it, affective experiences are subjective and vary with the decision maker. This marks a substantial difference between advice-taking in decision making and in revising factual opinions, and raises doubts regarding the usefulness of averaging affective forecasts.

To assess the impact of the (in)dependence between advisor's and advisee's prediction errors and the potentially different affective experiences that advisee and advisor are predicting on the usefulness of advice, consider the Normal model of pleasure levels and affective forecasts from Section 1.2.3. Let the advisee's pleasure levels be determined as $\mu_i^1 \sim N(\mu^1, \sigma^2_{\mu^1})$ and the advisor's pleasure levels as $\mu_i^2 \sim N(\mu^2, \sigma^2_{\mu^2})$. Assume that both advisee and advisor cannot predict (or recall, in the case of an advisor who has experienced an option previously) these pleasure levels perfectly, and allow for potential biases. Their affective valuations are then given by $V_i^1 \sim N(\mu_i^1 + b_i^1, \sigma^2_{V^1})$ and $V_i^2 \sim N(\mu_i^2 + b_i^2, \sigma^2_{V^2})$, respectively.

Consider now a decision maker who revises an initial affective forecast V_i^1 on the basis of the piece of advice V_i^2 according to Equation 1.1. Assuming that decision makers aim to minimise the probability of choosing suboptimally, taking others' advice then improves the outcome of a decision when choosing on the basis of the revised forecast V_i^{ad} decreases this probability compared to making the choice without advice, on the basis of the initial forecast only[10]. I will now discuss the

[10] A similar analysis can also be conducted on the basis of the expected disappointment. It yields similar results, so I will focus on the probability of choosing subopti-

effects of those three factors which jointly determine these probabilities, and hence the usefulness of advice: 1) the correlation $\rho_\mu = \rho(\mu_i^1, \mu_i^2)$ between the respective pleasure levels that advisee and advisor experience, 2) how accurate the advisor's affective valuations are in comparison to the advisee's initial affective forecast, captured by the relationship between $\sigma^2_{V^1}$ and $\sigma^2_{V^2}$, and 3) the correlation $\rho_V = \rho(V_i^1, V_i^2 | \mu_i^1, \mu_i^2)$ between the affective valuations of advisee and advisor, conditional on the pleasure levels.

1.3.3 Optimal advice-taking

The following analyses are normative and describe optimal behaviour. They capture how useful a piece of advice could *potentially* be, if advisees were to understand the characteristics of the choice situation as well as their relationship with the advisor perfectly. As such, they do not claim to describe the behaviour of a decision maker, but instead allow me to distinguish between choice situations in which others' advice is likely to be useful and those in which it likely is not. I report the optimal weights on the advisee's initial affective forecast and the advice as given by Equation 1.1 and computed on the basis of simulation methods.

1.3.4 Similarity in tastes

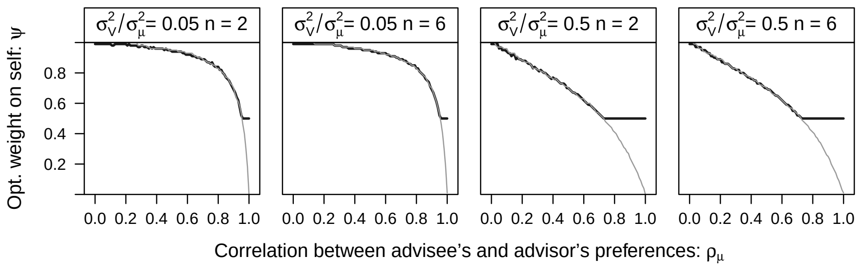
The first factor which determines the usefulness of advice is the correlation ρ_μ between the respective pleasure levels that advisee and advisor experience. It captures in how far advisee and advisor experience the same pleasure from a given option. If Claire and Pauline, for example,

mally as the measure for decision quality.

tend to like the same films, ρ_μ is large when they predict how enjoyable a new film will be. If, on the other hand, they sometimes agree but just as often disagree in how much they enjoy a film, ρ_μ is close to zero.

Consistent with the films example, let affective forecasts be imperfect but unbiased such that $b_i^1 = b_i^2 = 0$. Assume Claire is choosing from n new films which neither her nor Pauline have seen yet, and that they are equally good (or equally bad) at predicting their affective experiences on average, so that $\sigma^2_{V^1} = \sigma^2_{V^2}$. Furthermore, assume that the errors they make in their prediction are independent, so that $\rho_V = \rho(V_i^1, V_i^2 | \mu_i^1, \mu_i^2) = 0$. Finally, let $\mu^1 = \mu^2 = 0$ and $\sigma^2_{\mu^1} = \sigma^2_{\mu^2}$, an auxiliary assumption that will be maintained for all subsequent analyses[11]. Given these parameters, the black lines in Figure 1.5 show how the weight on the advisee's own affective forecast ψ which minimizes the probability of choosing suboptimally depends on the similarity between advisee and advisor ρ_μ, for choice sets of different sizes and with different variance ratios.

Figure 1.5: Advice-taking: Optimal weight on own affective forecast



The two panels on the left of Figure 1.5 show that when errors in

[11]This assumption can be interpreted as imposing that advisee and advisor are equally optimistic.

affective forecasts are small compared to the differences in pleasure associated with the alternatives, preferences have to be highly correlated for the advisee to place a substantial weight on the advisor's opinion. If the effect of the forecast errors is large, on the other hand, like in the panels on the right of Figure 1.5, this is not the case. For values of ρ between .75 and 1, advisees can place as much weight on the advisor's opinion as on their own ($\psi = .5$), and even when the correlation between advisor's and advisee's preferences is low, the weight they should place on the advice still remains substantial. Meanwhile, the smallest optimal weight ψ advisees should give their own affective forecasts is .5. The minimal ψ for which $\mathbb{P}(\mu_i^* \neq \mu^*)$ is smaller for the choice made on the basis of the advice $V_{i^*}^{ad}$ than for the choice made without the advice $V_{i^*}^1$ (shown in gray) can be lower than .5 for highly correlated preferences, but since advisees and advisors are assumed to be equally accurate, it is never optimal to place more weight on the advice than on one's own affective forecast. Finally, Figure 1.5 suggests that these effects are independent of the number of alternatives involved in the choice: there is no difference in the relationship between the minimal weight on one's own affective forecast ψ and the correlation of the preferences ρ between $n = 2$ and $n = 6$.

These findings have two implications. On the one hand, Figure 1.5 shows that the "‘Wisdom of the Crowd"’ indeed exists in decisions in matters of taste. Advice can help decision makers make more accurate affective forecasts and consequently make better decisions, even if the advisor is not more knowlegdable than the advisee. On the other hand, not any crowd will help. Different people experience the same situation differently, and advisors need to be similar to their advisees for the latter to benefit from their advice.

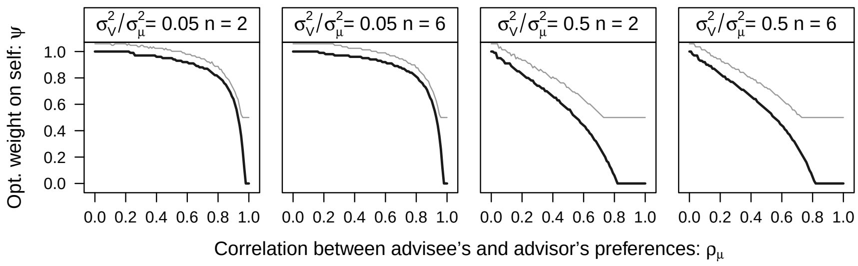
1.3.5 Accurate advice

Next, consider the role of accuracy. Some advisors are able to predict or report their affective experiences more accurately than their advisees. If Pauline were to give advice to a friend who knows less about films then she does, for example, her affective forecasts may be subject to smaller errors than her friend's. Pauline may have even seen a particular film already, and be able to recall how much she enjoyed it from memory. Let us begin by analysing decision environments which do not give rise to systematic biases so that $b_i^1 = b_i^2 = 0$. Accuracy is then determined by the variance ratio $\frac{\sigma_V^2}{\sigma_\mu^2}$, so that differences in accuracy can be analysed by comparing the variability of the prediction errors of the advisee, $\sigma_{V^1}^2$, and the advisor, $\sigma_{V^2}^2$.

How does the accuracy of the advice affect its usefulness for the decision maker? As before, let $\mu^1 = \mu^2 = 0$ and $\sigma_{\mu^1}^2 = \sigma_{\mu^2}^2$. For illustration purposes, consider the extreme case of an advisor who is perfectly accurate so that $\sigma_{V^2}^2 = 0$. The black lines in Figure 1.6 show the optimal weight ψ on the decision maker's own affective forecast, for choice sets of different sizes and with different variance ratios.

Figure 1.6 shows that the advice of perfectly accurate advisors should receive more weight than that of advisors who can predict their affective experiences only as accurately as their advisees from Figure 1.5 (shown in light gray for comparison). Again, the number of alternatives has no effect, and advice should be weighted more in choice settings where errors are large compared to those where errors are small. At a second look, Figure 1.6 reveals that if the advisor's preferences are sufficiently similar to the advisee's, the latter should make the decision *exclusively* on the basis of the advice. In all four panels in Figure 1.6, the curves, which show the *optimal* weight that the decision maker should place on the advice, are flat at $\psi = 0$ as ρ approaches 1. In this region, the decision maker should not place any weight on his own affective forecast, as the advice is a more accurate predictor of the affective experience. Furthermore, if the effect of the errors is large ($\frac{\sigma_V^2}{\sigma_\mu^2} = .5$), this region is sizeable: for all values of $\rho > .8$, the optimal weight to place on one's own affective forecast is 0. In other words, if my affective forecasts are subject to sufficiently large errors, I should follow the recommendation of an accurate advisor, even if the advisor's preferences do not perfectly mirror my own.

Figure 1.6: Optimal weight on own affective forecast for perfectly accurate advice



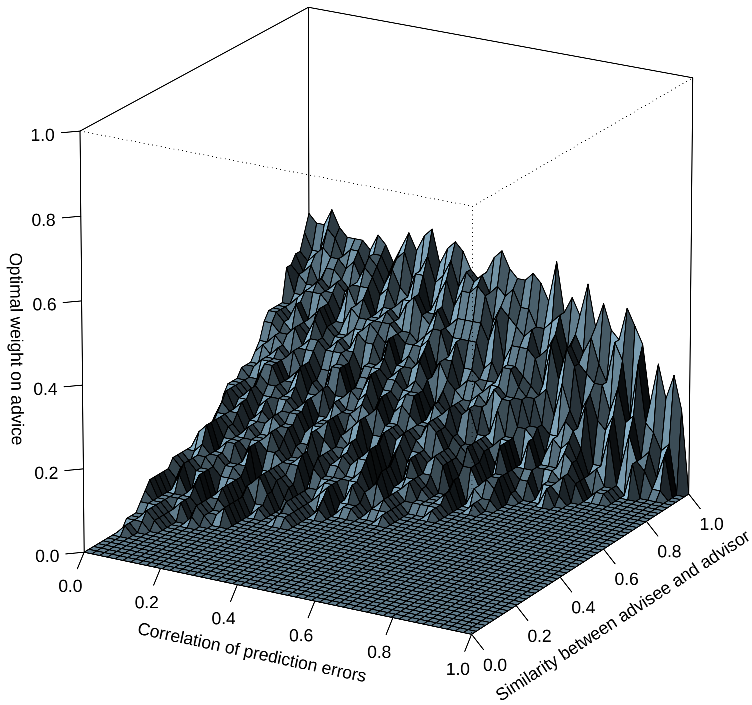
This finding does not only apply for advisors who are perfectly accurate, but holds for all advisors who are sufficiently more accurate than their advisees, which is important as perfectly accurate advisors rarely exist. Even experienced advisors, that is, advisors who have themselves already experienced the option(s) that the advisee is choosing from, can make mistakes when recalling their affective experiences: affective memory is not perfect and can even exhibit systematic biases as shown for temporally extended experiences by Kahneman et al. (1993). Nonetheless, experienced advisors will often be able to give accurate affective valuations, especially when the affective experience lies in the recent past.

1.3.6 Error correlation

The third factor which determines the usefulness of advice is how the errors that advisee and advisor make in forecasting (or recalling) their affective experiences depend on each other. It is captured by the conditional correlation $\rho_V = \rho(V_i^1, V_i^2 | \mu_i^1, \mu_i^2)$ between the affective forecasts of advisee and advisor, which can be thought of as the correlation betwen their prediction errors. To see how it affects the usefulness of advice, consider again the case in which advisee and advisor make equally accurate, unbiased affective valuations with $\sigma^2_{V^1} = \sigma^2_{V^2}$ and $b_i^1 = b_i^2 = 0$. Figure 1.7 shows how the correlation between prediction errors ρ_V and the similarity between advisor and advisee ρ_μ jointly determine the optimal weight on the advice $1 - \psi$ in a decision between $n = 2$ options with a variance ratio of $\frac{\sigma_V^2}{\sigma_\mu^2} = .5$.

The graph reveals that the weight the advisee should place on the advice increases as the correlation between the prediction errors ρ_V decreases. This is reminiscent of the "Wisdom of the Crowd" in estimation problems, where accuracy gains from averaging are larger if estimates are uncorrelated. Similarly, advice is more useful in decision making when the errors that advisee and advisor commit in their affective forecasts are independent. When is this likely to be the case? The discussion in Section 1.2 highlights the importance of uncertainties as a source for un-

Figure 1.7: Advice-taking: Optimal weight on advice



biased errors in affective forecasts. Uncertainties arise from the information a decision maker has access to when making a decision, and advisee and advisor are likely to experience different uncertainties if they have access to different information. An independent advisor who makes affective predictions based on a different information set than that of the advisee will therefore often give useful advice.

At the same time, Figure 1.7 also shows that the similarity between advisee and advisor is even more important than the independence of the advisor. The region in the lower left of the $(x;y)$ plane in the figure shows that if advisee and advisor are not similar to one another, advisees should not take the advice into account, even if their advisors' prediction errors are uncorrelated with their own. The top right region, on the other hand, shows that as long as advisee and advisor are very similar to one another, advisees should still place some weight on the advice, even if prediction errors are highly correlated. Furthermore, the results in Figure 1.7 are based on simulations with a variance ratio of $\frac{\sigma_V^2}{\sigma_\mu^2} = .5$, that is, an environment in which errors are relatively *large*. The effect is even more pronounced in environments with smaller variance ratios, in which the effect of errors is smaller. The symmetry of the model implies that only in environments where $\frac{\sigma_V^2}{\sigma_\mu^2} > 1$, that is, when the variability of the errors is larger than the variability of the pleasure levels, the independence of the advisor is more important than the similarity between advisor and advisee. Since such environments are few, as implied by the discussion in Section 1.2.3, these results confirm the importance of the observation that different people experience the same situation differently, and advisors need to be similar to their advisees for the latter to benefit from their advice.

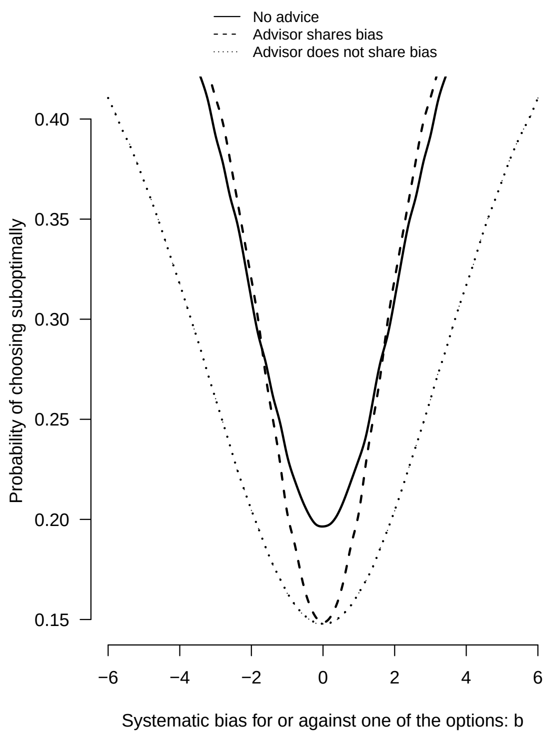
Finally, the results also strengthen the finding that advisees should place more weight on the advice of accurate advisors, and of advisors who can draw on previous experiences. First, think back to the extreme case of perfectly accurate advice. If the advisors do not make any errors in predicting their affective experiences, there cannot be a correlation between these and the errors of the advisee. Perfectly accurate advice is therefore useful advice, not only because of its accuracy but also because the (non-existant) prediction errors of the advisor cannot be correlated with the advisee's. An analysis of the more realistic case of an experienced advisor whose advice is relatively accurate but not perfectly accurate yields a similar conclusion. If prediction errors are unbiased, the advisors, who recall affective experiences from memory, are unlikely to make the same errors as the advisees, who predict their enjoyment of a novel experience. In other words, advisor and advisee draw on different psychological processes as well as different information sets, which results in a low correlation of their prediction errors. The same argument also applies for possible biases which affect advisor and advisee. Memory biases (Kahneman and Thaler, 2006) which may affect an experienced advisor, are likely to be unrelated to the biases which could affect the inexperienced decision maker.

1.3.7 Advice and biases

What are the more general implications of this analysis for advice-taking when affective forecasts are systematically biased? By definition, biases are related to the notion of predictive accuracy, as they represent the extent to which decision makers systematically mis-predict their affective reactions. The considerations about memory biases show that they are also related to the notion of correlated errors. To see how biases affect predictive accuracy and error correlation, consider a choice between two options where $\mu_1^1 = \mu_2^1 = \mu$. Further assume that the decision maker exhibits a bias b against one of the options so that $b_1^1 = b$ and $b_1^2 = 0$. The decision maker's affective forecasts are then given by $V_1^1 \sim N(\mu_1^1 + b, \sigma_{V^1}^2)$ and $V_2^1 \sim N(\mu_2^1, \sigma_{V^1}^2)$. For the purpose of the argument, assume that the advisor's preferences are perfectly aligned such that $\rho_\mu = 1$, and that the advisee incorporates the advice optimally with $\psi = .5$. Consider first the case in which decision maker and advisor are equally accurate in their predictions with $\sigma_{V^1}^2 = \sigma_{V^2}^2$ and share the bias:

$b_1^2 = b_1^1 = b$ and $b_2^2 = b_1^2 = 0$. Figure 1.8 shows the probability that the decision maker chooses suboptimally for different levels of bias.

Figure 1.8: Advice in a choice between two options with a bias ($\frac{\sigma_V^2}{\sigma_\mu^2} = .5$)



The solid black line in Figure 1.8 represents choices made individually, without advice. Its inverted U-shape implies that the larger the absolute value of the bias that the decision maker experiences, the larger the probability of choosing suboptimally (cf. Section 1.2.5). The same holds for a potential advisor: the larger the bias advisors experience, the less accurate their advice. The preceding discussion regarding advisor accuracy then implies that unsurprisingly, advisees should place less weight on the advice of advisors who are subject to biases in their affective predictions.

Next, compare the solid- to the dashed line, which depicts choices which incorporate the advice of an advisor who shares the advisee's bias. Figure 1.8 reveals that such shared biases reduce the usefulness of advice: even "perfect" advice based on perfectly aligned preferences and combined optimally with the affective forecast does not always lead to better choices. Only for values of b close to zero, the dashed line falls below the black line. Here, advice is useful, as it reduces the probability of choosing suboptimally. On the other hand, taking into account others' advice fails to improve decision quality when affective forecasts are largely determined by a shared bias. In this case, advice may even lead to worse decisions as it reinforces advisees in their initial -biased- affective forecast: the dashed line in Figure 1.8 is slightly above the black line if the absolute value of b is sufficiently large.

On the other hand, consider an advisor who does not share the advisee's bias, so that $b_1^2 = b_2^2 = 0$ with all other parameters unchanged. The dotted line in Figure 1.8 represents the probability of choosing suboptimally when incorporating the advice of such an advisor. For $b \neq 0$, the dotted line lies below both the dashed- and the black lines: taking into account the advice of an advisor who does not share the advisee's bias allows the latter to reduce the probability of a suboptimal choice. These findings are consistent with the previous analysis of the effect of correlated errors on the usefulness of advice. Shared biases will be reflected in a positive correlation between advisor's and advisee's predictions $\rho_V = \rho(V_i^1, V_i^2 | \mu_i^1, \mu_i^2)$, conditional on their respective pleasure levels. Advisees should therefore place more weight on the advice of advisors who do not share their biases.

When are advisees and advisors likely to suffer from the same biases, however? Think back to the example from Section 1.2.5 in which Claire suffers from diversification bias while simultaneously choosing two films which she plans to watch later the same day. Having decided that she wants to watch one film by a particular director, diversification bias makes her under-estimate the pleasure associated with watching a second film produced by the same director. The usefulness of advice in this decision is partly determined by whether Claire's advisor, Pauline, also experiences diversification bias. If Claire were to ask for her advice while in the store, Pauline would also have to make her affective predictions simultaneously for sequential consumption, and would likely suffer from the same bias. If, on the other hand, Pauline were to arrive home after Claire had already watched the first film, and Claire were to ask for her advice then, she would not be subject to the bias because she would make the prediction immediately preceding consumption.

The example illustrates how biases often arise from the choice environment. When making affective predictions simultaneously for sequential consumption, Pauline shares Claire's diversification bias. She shares the bias even if she has already seen the films herself: while she may be able to accurately recall how much she enjoyed each film from memory, the bias arises from the fact that the predictions are made simultaneously for sequential choice. When advisee and advisor share the same decision environment, they are likely to share biases, too. This reinforces the conclusion that advisors need to be independent for their advice to be useful, as independent advisors not only base their predictions on a different information set, but will also not fall prey to the same biases as the advisees.

1.3.8 Discussion

This section shows that when making decisions based on how much we expect to enjoy different options, we may be able to make better decisions if we take into account information about how much others would expect to enjoy them. The degree to which such advice will help depends on the similarity between our tastes and those of our advisors, and the nature and the size of the errors both we and our advisors commit in our affective forecasts. Advice is more useful when preferences are similar, and when it comes from accurate and independent advisors. This implies that advisors who have already experienced the options that the decision maker is choosing from will often give useful advice, since they are accurate and are less likely to share the errors and biases of the decision maker. The analysis also implies that even an inexperienced advisor can help us make better decisions.

The small existing experimental literature on advice-taking in choice situations finds behaviour consistent with these normative recommendations (Gilbert et al., 2009; Yaniv et al., 2011). Participants in Gilbert et al. (2009)'s experiments made more accurate affective forecasts of how they would experience (i) a speed date with a second participant and (ii) being subjected to a negative peer evaluation when told how much a second participant had enjoyed the same experience. Yaniv et al. (2011), on the other hand, find that people take into account the affective reports of an experienced advisor when making choices. In line with predictions from this analysis, the participants in Yaniv et al. (2011)'s experiments also give greater weight to others' advice if they perceive advisors as similar based on their behaviour or based on demographic factors.

It should also be noted that while Yaniv et al. (2011) use a choice setting in which participants' affective forecasts are unbiased, Gilbert et al. (2009) find that their participants on average underestimate how enjoyable a speed date would be, probably because of an attitudinal bias, and that they overestimate the negative effect of a negative peer evaluation because of impact bias. The present framework predicts that affective forecasts should be more accurate when they take into account the advice, both in the biased choice setting of Gilbert et al. (2009) and in the unbiased one of Yaniv et al. (2011). Only Gilbert et al. (2009) measure affective forecasts directly, but their findings confirm this hypothesis. Gilbert et al. (2009) do not analyse the similarity between participants, but since all their participants face the same choice setting which gives rise to a bias, their affective forecasts are largely determined by the bias. In such an environment, similarity is induced by the bias, and the advice of any experienced advisor who is not subject to the bias anymore will lead to more accurate affective forecasts.

Finally, I assume throughout this section that advisees know how similar their advisors' preferences are to their own. In the case of Claire and Pauline, this may be warranted: their friendship has taught them where their tastes are similar and where they differ. Friends and family are an important source of advice for our decisions, and I believe that the present results are applicable to advice-taking in a large number of decisions in which we can judge the similarity between us and our advisors accurately. With other advisors, however, we will be uncertain as to how similar their tastes are to ours, such as anonymous advisors on the Internet. How uncertainties about preference similarity and decision makers' beliefs regarding these uncertainties affect advice-taking is an interesting subject for future research.

1.4 General Discussion

This paper shows that making choices on the basis of predictions of future affective states can lead to suboptimal choices and post-decision disappointment and that taking into account others' advice can reduce these effects. Its three main conclusions can be summarised as follows. First, an interaction between the errors in affective forecasts and the choice process induces an asymmetry between positive and negative forecast errors. As a result, decision makers will on average experience disappointment even if forecast errors are unbiased, and biases in affective forecasts are more likely to have an effect on the outcome of a choice if they lead decision makers to over-estimate, rather than under-estimate how much they will enjoy an action. Second, advice can provide decision makers with additional affective predictions which can reduce the expected disappointment and the probability of choosing suboptimally, in the same manner that other people's estimates can help make better quantitative judgments. Unlike in judgment contexts, however, advisee's and advisor's preferences need to be similar for the advice to be useful. Third, when choosing advisors, decision makers should prefer those advisors who are similar to them, accurate and independent. Advisors who are not independent are likely to draw on the same information as their advisees and may even share their biases, which reduces the usefulness of their advice. Accuracy and independence are often found in people who have experienced the options that the decision maker is choosing from themselves previously, who consequently make for good advisors.

A number of important issues remain. The first concerns the magnitude of the disappointment incurred from maximisation based on imperfect affective forecasts. The framework developed here makes a general prediction that decision makers who maximise should on average be disappointed with the outcome of their choices. When they first discussed this phenomenon, Harrison and March (1984) went as far as saying that "a society that defines intelligent choice as normatively appropriate for individuals and organizations imposes a structural pressure toward postdecision unhappiness" (page 39). The question then becomes if we are really that unhappy with our decisions. I believe that this paper constitutes a first step towards answering this question by differentiating between different types of errors (unbiased errors vs. biases) and how they arise from different choice settings and environments, in the tradition of Brunswik (1956).

It is also possible, and consistent with the ideas of Harrison and March (1984), that some individuals do not give in to "the structured pressure toward unhappiness" by refusing to maximise and relying on different decisions rules instead. Schwartz et al. (2002), for example, show that some people show little maximising behaviour and aim to satisfy their aspiration levels instead. In a related study Iyengar et al. (2006) even find that people who maximise are often less happy with the outcomes of their choices than such "satisficers". This is consistent with an interpretation in which happiness is partly determined by post-decision surprise, as the present framework suggests that maximisers will experience more post-decision disappointment. Future research could investigate whether decision makers are more likely to satisfice rather than maximise when they are aware of the imperfections in their affective forecasts. Similarly, decision makers who are aware of the imperfections in their affective forecasts when making intertemporal decisions may resort to non-maximising decision rules. They could, for example, only commit as far as planning is possible and delay commitment to courses of action for which future affective reactions are hard to predict (Hogarth, 2010). Such decision rules could explain why decision makers are not constantly disappointed with the outcomes of the (intertemporal) decisions they make.

A related explanation is that decision makers are aware of the interaction between the errors they make in their affective forecasts and their choice processes and that they adjust their behaviour in order not to experience post-decision disappointment. This paper emphasizes advice as one method by which we can adjust our expectations and make better affective forecasts, but there are others. Highly rational decision makers could resort to Bayesian methods for adjusting their expectations downwards as suggested by Smith and Winkler (2006) in the context of decision analysis in organisations. Harrison and March (1984) suggest that decision makers who believe that their affective forecasts may be imperfect because their information about the alternatives involved in the choice is insufficient could aim to obtain more information, or to reduce the choice set.

As for advice in decision making, this paper identifies the similarity between advisor and advisee as a key factor in determining its usefulness. This observation gives rise to an important question: when and why are people similar to one another in terms of their tastes and goals?

It seems reasonable to think of similarity as a domain-specific characteristic. People may share a taste in music, for example, or agree on the criteria which make a job interesting and worthwhile. It is not as clear that liking similar music makes people also agree on whether a particular career path is better than an alternative one, although social influence may reinforce similarities across domains (DeMarzo et al., 2003). Such processes could be an interesting subject for future research, but common sense suggests that similarity will generally remain at least partly domain-specific. This explains why a high school graduate should probably consult with a friend, rather than his parents, when buying a new CD, but take into account his parents' advice when choosing what to study at university.

This chapter should also be a point of departure for further experimental work on advice in matters of taste. The role of experience in advice proposed in this paper, for example, has not been investigated yet, with existing experiments relying on experienced advisors (Gilbert et al., 2009; Yaniv et al., 2011). A second line of research should also explore advice-giving rather than advice-taking. One important question is what advisors communicate to advisees when left to their own devices. My analysis is concerned with advisors who communicate valuations of *their own* affective experiences, but what about advisors who communicate their forecasts of how much they would expect their advisees to enjoy an alternative? Empathy may then take over the role played by similarity in this paper, although some experimental evidence exists that shows that people are not good at predicting the preferences of proximate others (Davis et al., 1986).

Finally, research on advice in choice contexts should also be related to work on small groups. Claire and Pauline, for example, who have played an important role throughout this chapter, go to the cinema *together* whenever they can. Going to the cinema, a football match, or a shopping trip together with a friend is often more enjoyable because it allows us to share our impressions and experiences. A consequence of the framework presented in this paper is that it may also enable us to make better decisions, and experience less post-decision disappointment. Doing things together, and finding compromises may provide an effective means to take proximate others' opinions as seriously as the present research suggests we should.

The following chapter has previously been published as:

Müller-Trede, Johannes. (2011). Repeated Judgment Sampling: Boundaries. *Judgment and Decision Making*, 6 (4), 283-294.

It is available at http://journal.sjdm.org/11/101217a/jdm101217a.pdf

Chapter 2

REPEATED JUDGMENT SAMPLING: BOUNDARIES

2.1 Introduction

Imagine you have been asked to make a quantitative judgment, say, somebody wants to know when Shakespeare's Romeo and Juliet was first performed, or you might be planning a holiday in the Alps and are wondering about the elevation of Mont Blanc. An effective strategy to answer such questions is to make an estimate and average it with that of a second judge: a friend, a colleague or just about anybody else (see, for example, Stewart, 2001; Yaniv, 2004). What, though, if your colleague or friend is unavailable and cannot give you that second opinion? Recent research suggests that you could improve your answer by bringing yourself to make a second estimate and applying the averaging principle to your own two estimates (Herzog and Hertwig, 2009; Vul and Pashler, 2008).

The effectiveness of this suggestion, however, will depend on both the degree to which you are able to elicit two independent estimates from yourself and your willingness to average them. Previous research has focused on the method used to elicit the second estimate. The focus here lies on the type of question being asked, and its interaction with how successive estimates are generated. I report experimental results for different sets of questions which aim to be more representative of quantitative judgments (Brunswik, 1956). I first reproduce previous results which establish the existence of accuracy gains for year-estimation questions such as "In what year were bacteria discovered?" (Herzog and Hertwig, 2009). While I find similar gains for questions about percentage shares (e.g., "Which percentage of Spanish homes have access to the Internet?"), I do not find evidence of accuracy gains for general numerical questions such as "What is the distance in kilometers between Barcelona and the city of Hamburg, in Germany?" or "What is the average depth of the Mediterranean Sea?". I then investigate whether this difference can be explained by the degree to which answers to the various question types are implicitly bounded, but this hypothesis is not supported by the data.

A second factor is whether judges actually recognise the potential gains from averaging and behave accordingly. Larrick and Soll (2006) argue that people often do not understand the properties and benefits of averaging procedures. My experimental data provide further evidence: only a small minority of judges consistently average their estimates. Often, judges settle for one of their first two judgments as the final answer instead or even extrapolate, providing a final answer that lies outside of the range spanned by their first two estimates. They consequently fail to realise the potential gains from averaging.

2.1.1 Repeated Judgment Sampling

Efficiency gains from averaging are pervasive in different contexts and have been discussed extensively in the literatures on forecasting (Armstrong, 2001), opinion revision (Larrick and Soll, 2006) and group judgment (Gigone and Hastie, 1997). The phenomenon is well-understood: averaging leads to accuracy gains as long as the errors inherent in the estimates are at least partly independent (Surowiecki, 2004). Vul and Pashler (2008) and Herzog and Hertwig (2009), using different methods to sample multiple judgments from the same judge, found that averaging these also leads to accuracy gains.

In both of these studies, participants were not aware that they would have to answer the same question multiple times and were asked for their first judgment as a best guess. Vul and Pashler (2008) then simply asked the same person to make the same judgment again. They found an accuracy gain when the second judgment followed immediately, but reported a considerable increase in effectiveness if it was delayed for three weeks. Herzog and Hertwig (2009), on the other hand, proposed a method they called dialectical bootstrapping, which presents judges with instructions on how to make the second judgment, asking them to (i) re-consider their first judgment, (ii) analyse what could have been wrong, and specifically, whether it was likely too low or too high, and (iii) make a second estimate based on these considerations (p. 234). Using this method, they obtained larger accuracy gains than without instructions.

Finally, Rauhut and Lorenz (2011) used yet another method to elicit the judgments. In their experiment, participants had to provide five answers to the same question and they were informed about this at the outset. They confirmed Vul and Pashler's (2008) and Hertwig and Herzog's (2009) findings of positive accuracy gains from averaging two estimates for four of the six questions they analysed. Furthermore, they found that repeated judgment sampling had diminishing returns: accuracy gains decreased substantially when averaging more than two estimates from the same judge.

2.1.2 Process and environment

Vul and Pashler (2008) interpreted their initial finding as evidence for probabilistic representations of concepts in people's minds, but nobody has argued that the mechanism underlying repeated judgment sampling is the same as that leading to accuracy gains when averaging different judges' answers. So far, little is known about how judges generate their different judgments, although some suggestions have been made. Both Vul and Pashler (2008) and Herzog and Hertwig (2009) pointed out the possible role of anchoring-and-adjustment processes, and Rauhut and Lorenz (2011) conjectured that additional judgments may sometimes reflect people becoming emotional or talking themselves into taking wilder and wilder guesses.

A first step toward investigating the processes underlying repeated judgment sampling is to compare its performance in different environments. The experimental study reported below includes different types of questions, including a subset of the year-estimation questions used in Herzog and Hertwig (2009), percentage-share questions, and general numerical questions. I chose the latter two question types because they capture two common types of quantitative judgments judges could face in naturally occurring environments in accordance with representative design (Dhami et al., 2004). In addition, questions about percentage shares are on a response scale which is implicitly bounded between 0 and 100. This allows me to investigate whether the existence of such bounds affects the potential accuracy gains from repeated judgment sampling, as it has been shown to affect performance in other judgment tasks (Huttenlocher et al., 1991; Lee and Brown, 2004).

2.1.3 Potential and realised gains

A second issue is what judges actually do when asked to provide a third answer on the basis of their first two. This is an interesting question given people's reluctance to employ averaging strategies when combining their own opinion with somebody else's (Soll and Larrick, 2009), and neither Vul and Pashler (2008) nor Herzog and Hertwig (2009) asked judges to actually give a third estimate. In my analysis, I will distinguish between *potential* gains from averaging which I compute by taking the average of the judges' first two answers, and *realised* gains from their third and final estimates. Whether judges are more likely to average when both judgments are their own than when taking advice from somebody else is important for anyone who thinks of using repeated judgment sampling in actual decisions. In addition, how judges manipulate their previous answers in order to arrive at a third one may enable us to infer something about the processes that underlie the generation of estimates.

2.2 Experimental method and results

I report the results of an experimental study based on a judgment task with two stages. The first stage assesses repeated judgment sampling's performance in the context of different types of questions. It includes three different question types (within-subject) and either provides explicit bounds for the judges or does not (Bounds vs. No-bounds conditions, between-subject). In the second stage, judges are asked to provide a final estimate on the basis of their first two estimates (Self condition). Judges in a control condition are also given the two answers of a different judge, chosen at random from the participants of the experiment (Other condition). Participants were 82 undergraduate students from the subject pool of the Laboratory for Economic Experiments at Universitat Pompeu Fabra, Barcelona. They received an average payment of 8.70 Euro based on the accuracy (median percentage error) of their answers. Participants came from 16 different academic fields of study, and 58% were female.

2.2.1 Part I: Question type

The first part of the experiment analyses the effect of the question type on potential accuracy gains from repeated judgment sampling. All gains discussed in this section are like those reported in Herzog and Hertwig (2009), computed by taking the average of participants' two estimates, and comparing this average to their first answer. They are not "real" gains, since judges were not asked to provide a third answer themselves until the second part of the experiment. The results reported in this section aim to answer the question whether judges could potentially benefit from the method in different environments.

Method All participants first answered three blocks of twenty questions each (shown in the Appendix). The first block included a subsample of the year-estimation questions used in Herzog and Hertwig (2009). It was followed by questions about percentage shares, and the final set of questions consisted of twenty general numerical questions, the answers to which vary by many orders of magnitude. General numerical and percentage share questions were general-knowledge questions, partly sampled from local newspapers. After completing an unrelated choice task, all participants had to answer the same questions again, in the same order. The elicitation method was adopted from Herzog and Hertwig (2009), and used "consider-the-opposite"-type instructions as described above. To further ensure comparability, I also adopted their payment scheme and participants were paid on the basis of the more accurate of the two answers.

Throughout the experiment, subjects in the Bounds condition were also given explicit lower and upper bounds for their answer with each question. For year-estimation items they were told the answer was between 1500 and 1900 and for percentages between 0 and 100. For general numerical questions, the ranges depended on the true unknown value.[1] Subjects in the No-bounds condition did not receive this additional information.

The data were screened for anomalies before the calculations. The answers of eight participants, five from the Bounds condition and three

[1]See Appendix B; bounds were constructed so that the distribution of true values with respect to the bounds resembled those of the other two categories.

from the No-bounds condition, were dropped because they were missing a substantial number of answers. The analyses reported below are based on the answers of the remaining 74 (bounds: 28, no-bounds: 46) participants.

Results Because the distributions of the answers were skewed, the data were transformed to logarithms. Despite this normalisation, the *size* of the effect depends on the response range for each question. Since these differ considerably across question types, I refrain from estimating general models which include a variable for the question type and its interaction with the condition (Bounds vs. No-bounds). Instead, I compute separate regressions according to Equation 2.1 for each of the three question types.

$$y_{iq} = \alpha + \beta b_i + \delta_i + \theta_q + \epsilon_{iq} \tag{2.1}$$

Equation 2.1 describes a linear regression model with crossed random effects. In this framework, y_{iq} denotes the dependent variable (for the ith individual on the qth question), α is the main effect for gains, β the effect of the explicit bounds provided in the Bounds condition, and δ_i and θ_q denote random effects for individuals and questions, respectively. For each of the three question sets, I estimate five such regressions using different dependent variables, measuring the accuracy of the judges' two estimates and the potential gains judges could obtain from averaging their answers. All of these measures are based on the logarithms of mean absolute deviations of the various estimates from the true value; their algebraic formulae are presented in Table 2.1 (overleaf).

In Table 2.1, $x_{1,iq}$ and $x_{2,iq}$ refer to the first and second estimates of judge i for question q, respectively, and x_{tq} refers to the true value for that question. The first two entries in Table 2.1 are simply logarithms of absolute deviations from the true value.

Table 2.1: Measures of accuracy and accuracy gain.

Explanation	Formula
Accuracy of 1st est.	$MAD_{1st} = \left\| ln(\frac{x_{1,iq}}{x_{tq}}) \right\|$
Accuracy of 2nd est.	$MAD_{2nd} = \left\| ln(\frac{x_{2,iq}}{x_{tq}}) \right\|$
Gain: Rep. J. Sampling	$G_{RJS} = MAD_{1st} - \left\| ln(\frac{\sqrt{x_{1,iq}x_{2,iq}}}{x_{tq}}) \right\|$
Dyadic Gain	$G_{Dyad} = MAD_{1st} - \frac{\sum_{j \neq i} \left\| ln(\frac{\sqrt{x_{1,iq}x_{1,jq}}}{x_{tq}}) \right\|}{N-1}$
Gain: overall average	$G_{WoC} = MAD_{1st} - \left\| ln(\frac{(\prod_i x_{1,iq})^{\frac{1}{N}}}{x_{tq}}) \right\|$

The bottom three rows in Table 2.1 describe the different measures for accuracy gains. All three are computed as simple differences in absolute value with respect to the error of the first estimate. A positive coefficient therefore implies an accuracy gain over the first estimate.[2] Second, they are all based on geometric means because of the skew of the answers. For repeated judgment sampling (G_{RJS}) the geometric mean is

[2] One could also define analogous measures for accuracy gains with respect to the second estimate. Since there is no difference in accuracy between the two estimates (see Table 2.2 below), however, I chose to conduct the analyses in comparison to the first estimate only.

simply the square root of the product of a judge's two estimates. "Dyadic gains" (G_{Dyad}) can be thought of as the expected accuracy gains from averaging with the estimate of a second participant drawn at random. They are computed as the average of the geometric mean of a judge's first estimate with the first estimate of a second judge. Finally, the estimate from averaging with all other judges at once—the "Wisdom-of-Crowds gain" (G_{WoC}) —is calculated on the basis of the geometric mean across all participants' first answers. It reflects the accuracy gain a participant could achieve by replacing his own estimate by the (geometric) mean of all participants' estimates.

Table 2.2 summarises the results of the analysis. For all three question types, and for both conditions, it provides coefficient estimates for the various accuracy measures discussed. All coefficients reported in Table 2.2 are significantly different from zero at the one per cent level except for the coefficient for accuracy gains from repeated judgment sampling for general numerical questions (marked by a dagger†), which is not statistically significant.

The results in Table 2.2 (overleaf) suggest that repeated judgment sampling may not lead to accuracy gains for all types of questions. The first two columns replicate Herzog and Hertwig's (2009) findings: repeated judgment sampling leads to accuracy gains for year-estimation questions, albeit smaller ones than those which can be expected from averaging one's estimate with that of another judge, or other judges. These results are confirmed for questions about percentage shares, and the effect is of similar size: accuracy gains from repeated judgment sampling are between a quarter and a third of the size of Dyadic gains, and between an eighth and a tenth of the size of the accuracy gains obtained

Table 2.2: Accuracy and potential gains by question type and condition.

	Year-Estimation		Percentage		Numerical	
	No-bnds	Bnds	No-bnds	Bnds	No-bnds	Bnds
MAD_1	.09	.07	.65	*.65*	1.8	.50
MAD_2	.09	.06	.63	*.63*	1.8	.51
G_{RJS}	.003	*.003*	.03	*.03*	.01†	*.01*†
G_{Dyad}	.008	*.008*	.10	*.10*	.24	.05
G_{WoC}	.019	*.019*	.23	*.23*	.61	.16

† All coefficient estimates are significantly different from 0 at the 1% level, apart from the one marked by the dagger which is not significantly different from 0. A coefficient in italics in the bounds condition indicates the absence of a treatment effect, resulting in the same coefficient estimate as in the no-bounds condition.

from averaging all participants' estimates. For general numerical questions, on the other hand, the picture is different. Averaging with other judges' answers improves accuracy, but there is no evidence of accuracy gains from repeated judgment sampling for these questions. The coefficient estimate for G_{RJS} is .01, which is 24 times smaller than the estimated coefficient for Dyadic gains and is not significantly different from 0 (p=.67).

Next, consider the effect of the bounds. I hypothesized that the difference between year-estimation, percentage share, and general numerical questions was the degree to which answers to these questions were implicitly bounded. The spectrum ranged from percentage share questions with their implicit bounds between 0 and 100 to general numerical questions, which had no obvious bounds associated with them. Year-estimation questions can be thought of as in between the two extremes, given judges' familiarity with the Gregorian calendar. The results in

Table 2.2 suggest that the provision of bounds indeed affects judges' performance differently depending on the question type. They do not support the hypothesis that bounds on the range of possible answers are a sufficient condition for the existence of accuracy gains from repeated judgment sampling, however. As hypothesized, judges' performance on percentage share questions is not affected by the provision of bounds at all. Bounds slightly improve accuracy for year-estimation questions, but do not effect the potential accuracy gains from the different averaging methods. They have a stronger effect on general numerical questions, with a more pronounced improvement in terms of accuracy, and effects on both Dyadic and Wisdom-of-Crowds gains. Note that these latter effects are negative: bounds reduce the accuracy gain which can be expected from averaging (although first answers are more accurate when bounds are provided, so while the improvement is smaller, it is an improvement over a more accurate first answer). Potential accuracy gains from repeated judgment sampling, on the other hand, are not affected either positively or negatively by the provision of bounds.

2.2.2 Part II: Third estimates

Method Having completed the first part of the experiment, all participants were asked to make a final judgment for a subset of fifteen questions, five for each question type. Participants in the treatment or Self condition had to make this estimate on the basis of their previous two estimates, while these were displayed on screen. The exact instructions they were given were the following: "For the last time, we would like to present you some of the questions which you have answered during this experiment. On the basis of your previous responses, we would like to ask you for a third answer. For this part of the experiment, you will be paid up to 8 Euros, based only on the accuracy of this third and final answer"[3]. The wording of the instructions was chosen so that participants would have no reason to believe that the subset of 15 answers was selected depending on the accuracy of their previous estimates, and to make clear that only accuracy mattered. In order to avoid priming subjects in a mindset which would make them average less, they were told to give the final answer "on the basis of their previous answers".

Participants in a control condition (Other) had to make the final judgment on the basis of their own two answers as well as the two answers of a different judge chosen at random among the other participants of the experiment. They did not have any information regarding the order of the two judgments from the second judge. Their instructions were similar: "For the last time, we would like to present you some of the questions which you have answered during this experiment. Here, you can see both your own two previous answers and the two answers of another participant of this experiment, who has been chosen at random. On the basis of this information, we would like to ask you for a third answer. For this part of the experiment, you will be paid up to 8 Euros, based only on the accuracy of this third and final answer."

Of the 82 participants in the experiment, seven were missing a substantial number of answers and had to be dropped. The analyses reported below are on the basis of the answers of 52 participants in the Self condition and 23 participants in the Other condition. Because of software issues, answers to the last question that was asked were not recorded correctly for a large number of participants, so it was also ex-

[3]Original instructions were in Spanish (written and translated by the author).

cluded from the analysis, restricting the latter to five year-estimation, five percentage-share and four general numerical questions.

Results The data from this part of the experiment can be used to answer two questions: How do judges arrive at their third answer?, and: Are third answers actually more accurate than first answers, as repeated judgment sampling suggests? To preview the findings of the analysis, different judges arrive at their final answers differently, but only a small minority of judges average consistently. Final answers are not significantly more accurate than first (or second) answers, and judges do not realise the potential gains from repeated judgment sampling.

As a starting point for the analyses, assume that judges in the Self condition arrive at their third judgment by taking a weighted average of their first two estimates. Denoting by ψ the weight placed on the first estimate, their final estimates can then be expressed as in Equation 2.2, a framework adopted from the literature on opinion revision (Soll and Larrick, 2009):

$$x_3 = \psi x_1 + (1 - \psi) x_2 \tag{2.2}$$

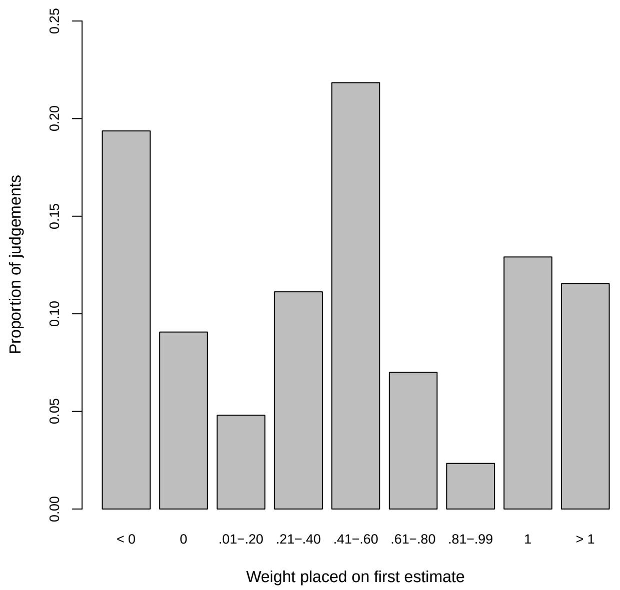
The value of ψ can then be calculated separately for each final answer. Note that this method cannot be applied to judges in the Other condition, where the corresponding expression is an equation in three unknowns.[4] The Other condition was included as a standard of com-

[4]Equation 2.2 is derived from $x_3 = \psi_1 x_1 + \psi_2 x_2$, under the assumption that $\psi_1 + \psi_2 = 1$. Judges in the Other condition had access to four pieces of information, so that their final answer should be a function of all four of them: $x_3 = \psi_1 x_{1,self} + \psi_2 x_{2,self} + \psi_3 x_{1,Other} + \psi_4 x_{2,Other}$. The assumption $\psi_1 + \psi_2 + \psi_3 + \psi_4 = 1$ is not sufficient to be able to calculate these weights for each item separately.

parison for the gains judges realise.

Figure 2.1 shows the distribution of ψ, aggregated over both questions and participants. From the figure, two assertions can be made about judges' behaviour.

Figure 2.1: Aggregate distribution of weight on the first estimate.



The first observation is that judges often extrapolate and provide a final answer outside of the range spanned by their first two answers, as indicated by the left- and rightmost columns in Figure 2.1. This constitutes a marked difference from the literature on advice-taking and opinion revision, in which estimates outside of the bounds spanned by one's own estimate and that of the advisor tend to account for less than 5% of answers (Soll and Larrick, 2009; Yaniv and Kleinberger, 2000). In comparison, in the present study such answers account for over 30% of all answers. While in opinion revision, they may be attributed to error and hence be disregarded (Yaniv and Kleinberger, 2000), it seems hard to make such an argument in the present case.

A second observation concerns the skew to the left evident in Figure 2.1. Judges tend to lean more toward their second answer than their first when giving a final answer: 44% of the aggregated judgments lie to the left of the central column in Figure 2.1, compared to only 33% to its right. This effect is not as strong as the Self/Other effect in advice-taking (Yaniv, 2004), but in the present context, both answers are one's own. The skew can also be detected in judges' behaviour at the individual level. Comparing the number of questions on which a particular judge uses weights with $\psi < .4$ with the number of questions on which he uses weights with $\psi > .6$, 60% of judges lean more toward their second estimate, and only 33% lean more toward their first.

What else can be said about how individual judges arrive at their final answers? Do they all behave similarly or are there individual differences? In particular, are there judges who consistently average their first two answers? In order to answer these questions, I compute a second measure which is closely related to ψ. For each answer, I calculate how far a judge deviates from taking an average:

$$\psi_d = |\psi - .5|$$

A reliability analysis shows that ψ_d is a reliable measure of individual differences, with a standardised Cronbach's alpha of .85. For each judge, I then calculate the median[5] ψ_d across the 14 questions, the distribution

[5]Possible values of ψ_d range from 0 to infinity, so the median seems to be the more

of which is shown in Figure 2.2. This median characterises judges in terms of how far they deviate from averaging. A median ψ_d smaller than .1 implies that a judge averages on at least 50% of answers; a median larger than .5 implies extrapolation for at least 50% of answers.

Figure 2.2: Distribution of judges' tendency to average.

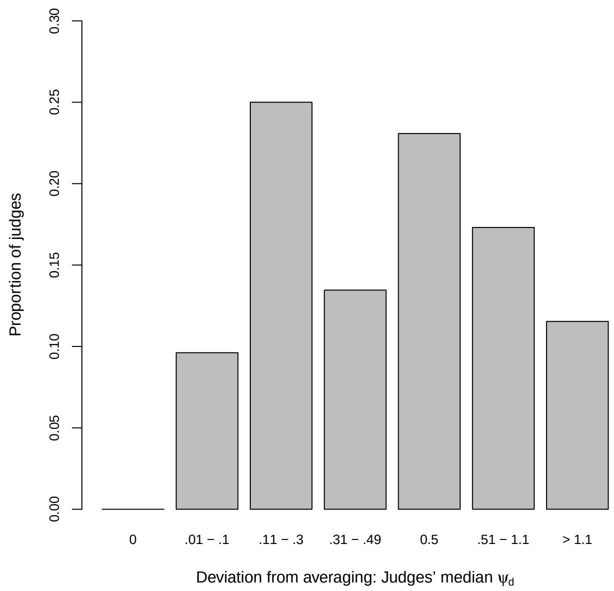


Figure 2.2 shows that around 10% of judges average consistently, resulting in a median ψ_d lower than or equal to .1. It also confirms the importance of extrapolations: 28% of judges exhibit a median ψ_d larger than .5, and therefore extrapolate on more than half of the 14 questions, providing final answers which lie outside of the bounds spanned by their first two estimates. Finally, for almost 25% of judges, the median ψ_d is

sensible measure of central tendency than the mean.

exactly .5. This does not imply that 25% of judges consistently settle for either of their first answers as their finale estimate, however, as this figure also includes judges who mix strategies and in addition to sometimes providing one of their previous answers as their third answer, average roughly as often as they extrapolate.

What are the implications for the actual accuracy gains judges were able to realise when making their final estimates? Table 2.3 shows both the *potential* gains from averaging, computed as before by taking the average of the judges' two estimates, and the *realised* gains, that is, the accuracy gain of the third and final answer over the first answer. Since I show above that there are no potential gains for general numerical questions, I conduct this analysis on the basis of the 10 year-estimation and percentage-share questions only.

Table 2.3: Realised- and optimal gains by condition

Condition	MAD_{Final}	Realised gain	Potential gain
Self	$.32^{***}$	.008	$.03^{**}$
Other	$.30^{***}$	.027	$-.20$

Significance levels: *** 1%, ** 5% level; H_0: coeff. estimate = 0.

Table 2.3 shows that judges in the Self condition were unable to realise significant accuracy gains. Potential gains, on the other hand, are positive for judges in the Self condition, who could have improved their judgment accuracy by simply averaging their previous estimates. Judges in the Other condition were not able to realise any gains, either, but unlike judges in the Self condition, they would not have reliably benefited from averaging. The coefficient estimate for potential gains is estimated at −.2, and is not statistically different from zero.[6] This explains why judges in the Other condition are not significantly more accurate in their final judgments than judges in the Self condition as can be seen in the first column of Table 2.3.

Finally, do realised gains differ between individuals? Maybe judges who average consistently improve in accuracy, while those who extrapolate do not. To answer this question, I correlate the judges' median ψ_d with the average gains they were able to realise for the 10 questions. If judges who average consistently outperform their fellow participants, this correlation should be significantly negative. The analysis does not yield significant evidence that "averagers" do better, however: Spearman's rank correlation coefficient is estimated at −.13 (p=.34). A more complex analysis could aim to answer the question at the more disaggregated level of individual answers instead, but a regression analysis with crossed random effects finds no significant effect of ψ_d on the final accuracy gain achieved on a particular question, either. The estimate for the coefficient associated with ψ_d is −.01 and fails to reach significance (p=.45).

[6]Note that this finding does not contradict the results in Part I according to which potential gains from dyadic averaging are *on average* higher than those from repeated judgment sampling. In Part I, gains from dyadic averaging are expected values over all other participants in the experiment; here, judges were paired at random with another participant. The pairings were such that potential gains were not significant. This in itself is an interesting observation that suggests that gains from repeated judgment sampling may be less variable than gains from dyadic averaging. In terms of comparing realised- and potential gains, however, it defeats the purpose of using the results from the Other condition as a comparative standard for those from the Self condition.

2.3 Discussion

In this paper, I provided new evidence that sampling more than one judgment from the same judge and averaging them can lead to accuracy gains in judgment tasks. For a sub-sample of the questions used in Herzog and Hertwig (2009), which ask judges to estimate the year in which a particular event happened, I replicated their finding of potential gains from repeated judgment sampling. I then confirmed this result for a second set of questions in which judges estimate percentage shares. On the other hand, I showed that repeated judgment sampling does not lead to accuracy gains for a third set of general numerical questions. Finally, I reported experimental data on how judges combine their two estimates when asked to do so, a question not previously addressed in the literature. The majority of judges did not consistently average their answers. In the experiment, they failed to realise the potential accuracy gains from repeated judgment sampling.

The finding that accuracy gains from repeated judgment sampling depend on the question being asked constitutes a challenge to the analogy drawn by Vul and Pashler (2008) and Herzog and Hertwig (2009) between repeated judgment sampling and the "Wisdom of Crowds". Accuracy gains from repeated judgment sampling behave like those from averaging different people's estimates for two of the three question sets I examine, but not for the third set of questions. While the source of the accuracy gains in repeated judgment sampling –the averaging principle– is doubtlessly the same as when averaging with somebody else's estimate, how judges generate their successive estimates remains unclear.

While my data fall short of answering this question, they reveal cues about what might be going on in the judges' minds. When asked for a third answer, judges often exhibit a reluctance toward averaging their first two answers, and many of them extrapolate outside of the range spanned by their first two answers. This suggests that they may have thought of more information which could be relevant for the question and which they had not considered when giving their previous estimates. Successive answers could then reflect how judges mentally integrate this cumulative information retrieved from memory to make their judgment. This account of repeated judgment sampling is also consistent with the findings that third estimates lean more toward the second, rather than the first estimates, and that the method does not always emulate the "Wisdom of Crowds". If the variability in the estimates is caused by different pieces of information, judges need at least some knowledge about a question for them to be able to benefit from repeated judgment sampling. On the other hand, even an ignorant judge would benefit from the "Wisdom of Crowds".

This notion is closely related to Rauhut and Lorenz's (2011) hypothesis that question difficulty affects potential accuracy gains, as it predicts no accuracy gains for a hard question that a judge does not know enough about. An interesting issue is what would happen for easy questions judges know a lot about, as these could include professional or expert judgments. Would experts benefit from repeated judgment sampling? In this context, note that my findings also qualify Rauhut and Lorenz's (2011) result that sampling more than two opinions from the same judge is subject to strongly diminishing returns, since all their questions are of the general-numerical type, which are here shown to be the type of questions repeated judgment sampling performs worst on. It is conceivable that, for easy questions, accuracy gains are particularly large, and that returns from sampling more than twice diminish more slowly.

A final consideration concerns the role of the instructions. On the one hand, the effects of the "consider-the-opposite" technique (Herzog and Hertwig, 2009), designed to induce judges to give two independent estimates could have persisted longer than intended and influenced final answers in the present experimental setup. This could have contributed to the judges' relutance to average their answers, and also to their tendency to extrapolate. On the other, the finding that only a relatively small minority of judges average consistently has implications for the instructions that would have to be provided, were repeated judgment sampling to be used in decision support. Since judges do not average voluntarily, for the technique to be effective, somebody has to average their judgments for them. That judges should be aware of this when asked for their judgments seems reasonable, even inevitable if a judge were to use the technique more than once. Future work should therefore examine the effects of informing judges about the benefits of averaging *before* eliciting their judgments.

Appendix: Questions used in the experiment and their associated bounds

Tables 2.4 to 2.8 display all sixty questions which were used in the experiment, translated from Spanish, and their respective answers. Some of the questions have been abbreviated slightly for formatting purposes; the published version of the paper features more accurate translations. The tables also include the bounds provided to subjects in the Bounds condition. The questions for which judges had to provide third estimates in the second part of the experiment are indicated with daggers†.

The bounds for the general numerical questions were constructed so that in absolute distance to the closest bounds, the distribution of the true values would resemble those of the other two question types. The mean absolute distance to the closest bound, as a percentage of the distance between the lower- and the upper bound is 0.27 for year-estimation, 0.25 for percentage-share, and 0.28 for general numerical questions. The associated standard deviations are 0.16, 0.13 and 0.13, respectively.

Table 2.4: Year-estimation questions: In what year...

Question	Ans.	Bounds	
		Low	High
... was the university of Harvard in Cambridge, MA (USA) founded?	1636	1500	1900
... was the first pocket watch built?	1510	1500	1900
... was the grammophone invented?	1887	1500	1900
... did works on the Palace of Versailles begin?†	1661	1500	1900
... were bacteria discovered?	1676	1500	1900
... did Benjamin Franklin invent the lightning conductor?	1752	1500	1900
... was barbed wire patented?	1875	1500	1900
... did the plague hit the city of London?	1665	1500	1900
... was electricity discovered?†	1733	1500	1900
... were the concerts for violin 'The 4 Seasons' published?†	1725	1500	1900
... was the thermometer invented?	1592	1500	1900
... was the first fan produced?	1711	1500	1900
... was dynamite invented?†	1866	1500	1900
... did the religious wars begin in France?	1562	1500	1900
... did the English fleet destroy the Spanish Armada?	1588	1500	1900
... was the last woman murdered for witchery in Europe?†	1782	1500	1900
... was Shakespeare's 'Romeo and Juliet' premiered?	1595	1500	1900
... was the English Bill of Rights passed?	1689	1500	1900
... was the Braille scripture invented?	1825	1500	1900
... was the first public screening of a film?	1895	1500	1900

Table 2.5: Percentage-share questions: Which (is the) percentage...

Question	Ans.	Bounds Low	High
... of the adult population in Spain who smoke on a daily basis?†	27	0	100
... of the Masters students in the Master in Economics at Universitat Pompeu Fabra in 2010 who are foreigners?	85	0	100
... of votes in Catalunya that CiU obtained in the last general elections?	21	0	100
... of its annual income that an average household in Spain spends on alcohol and tobacco?	3	0	100
... of Spanish homes that have access to the Internet?	54	0	100
... of the adult population in Spain that has completed third-level studies?	29	0	100
... of world GDP comes from the USA and the EU combined?	55	0	100
... of the population of Spain that lives in Catalunya?	16	0	100
... of Internet users connect from China?†	21	0	100
... of the 159 'Clasicos' played in the Spanish league that FC Barcelona has won?†	48	0	100
... of the people who live in Barcelona are 65 years old or older?	20	0	100

Table 2.6: Percentage-share questions: Which (is the) percentage...

Question	Ans.	Bounds	
		Low	High
... of the Spanish population that earned 6000 Euros or less in 2009?†	23	0	100
... of the Spanish population who would prefer to have a business of their own to being a employee, if they had sufficient resources?	40	0	100
... of civil servants who went on strike in the general strike on June 8th 2010, according to the government?	11	0	100
... of clients who change their tele-com provider do so primarily to save money?	75	0	100
... of women working in Catalunya who are in executive positions?†	7	0	100
... of the time they spend on-line do Spanish Internet users dedicate to social networks?	20	0	100
... of Spanish women who have suffered from domestic violence at least once in their lives?	25	0	100
... of employees in Spain who knew with certainty that they would lose their jobs during the next six months in June 2010?	13	0	100
... of the adult population in Spain who call their mum at least once when on holidays?	40	0	100

Table 2.7: General numerical questions: How many / What is ...

Question	Ans.	Bounds	
		Low	High
... underage homeless did the Generalitat have to support in 2009?†	1481	1000	3000
... ZARA stores are there in the city of Barcelona?	12	0	100
... the height of the highest elevation in Montseny, in metres?	1712	1000	3000
... victims (injuries and deaths) did the terror attacks on the Madrid Metro claim in 2004?	2049	1000	3000
... million Euros did FC Barcelona pay for new players in the season 2009/2010?	101.5	50	200
... minors between 14 and 17 were detained for drug-use on the street in Barcelona in 2008 and 2009?	1323	1000	3000
... modern Summer Olympics have been celebrated?	29	0	100
... Spanish soldiers are currently deployed in oversea missions?	2600	1000	3000
... Euros of public investment did the 2010 Pressupost of the Generalitat provide for?	$6*10^9$	$5*10^9$	$9*10^9$
... the distance between Barcelona and the city of Hamburg in Germany in kilometres?	1815	1000	3000

Table 2.8: General numerical questions: How many / What is ...

Question	Ans.	Bounds Low	Bounds High
... calories is the recommended daily intake of an adult woman?	2000	1000	3000
... the life expectancy of a baby born in Spain in 2009?†	80	0	100
... homes will be built in Catalunya with financial support of the Spanish central government in 2012?	1850	1000	3000
... 'municipios' are there in Catalunya?	946	0	1000
... the population of Barcelona in millions?	1.62	0	3
... days of rain are there in Barcelona each year on average?	72	0	100
... the average depth of the Mediterranean Sea, in metres?†	1500	1000	3000
... is the speed of sound, in kilometres per hour?	1236	1000	3000
... million passengers flew in and out of Barcelona airport in 2009?†	30	0	40
... Catalunya's GDP per capita in 2008, in Euros?†	29757	0	$3*10^5$

Chapter 3

ON COMPARING INDIVIDUAL AND GROUP BEHAVIOUR IN STRATEGIC DECISIONS

3.1 Introduction

The study of group decisions has a long history in the social sciences, and continues to be relevant in a world where many important decisions in economic, political and bureaucratic organisations are made by groups, teams, or committees, and not by single individuals. This includes strategic or interactive decisions, of course, in which outcomes depend on the choices of more than one decision maker. In this article, I review the findings of a research programme in behavioural game theory which has established differences between the behaviour of individual actors and that of small groups in a number of such strategic decision tasks. Groups have been found to give smaller amounts to one another than individuals in Dictator games (Luhan et al., 2009) and coordinate better with each other than individuals in a number of coordination games (Feri et al., 2010), for example. Some researchers have interpreted these findings as evidence of groups being able to make 'better' or 'more rational' strategic decisions, but few studies offer theoretical explanations as to why this should be the case. In this essay, I address this question by proposing a theoretical framework which models a number of general effects which are likely to make group behaviour differ from individual behaviour. I then re-assess the previous results from the literature in the light of this framework and find that it is still early to conclude that groups are better strategic decision makers.

This essay aims to be useful to behavioural scientists interested in group decision making whether they are working in social psychology, in organisational behaviour or in experimental economics. For experimental economists, it provides a review of what is known about group behaviour in experimental games as well as an alternative explanation for the finding that groups coordinate better than individuals. At the same time, it features an overview of the different decision tasks or games in which group behaviour has been examined for researchers in social psychology and organisational behaviour who are interested in group behaviour but less familiar with behavioural game theory (see Section 3.2, *The Data*). It also identifies some methodological advantages and disadvantages of using the different games in behavioural research, as well as a framework which helps to think about the main experimental manipulation commonly employed in these experiments (see Section 3.3, *The Individual vs. Group Paradigm*).

The framework proposed here views groups as decision environments in which multiple individual decision makers interact, and which then aggregates their preferred courses of action into a group action. I identify three dimensions on which decision making in a group setting differs from individual decision making. One, in a group setting, individuals can update their beliefs on the basis of the behaviour of their fellow group members before making a decision. Two, the outcome of a group decision often affects more people than that of a comparable decision made individually. Finally, three, group behaviour is the result of an aggregation process. I focus on these three effects because they are relatively general and are likely to affect group behaviour in many different decision tasks.

Conceptually, my analysis draws on Steiner's framework of group productivity which I adapt to the context of strategic group choice. Steiner (1966, 1972) distinguishes between task demands, resources and process as the three main determinants of group productivity. Similarly, I frame group choice as the result of a process in which the beliefs and preferences (resources) of individual group members in a given decision task (task demand) first interact with each other and are then aggregated by a specific decision rule (process). I employ Schelling (1980)'s classification of strategic decision tasks or games (see also Colman, 1982) to distinguish between (pure) coordination, mixed-motive, and constant-sum games. Coordination games are games where players' interests are perfectly aligned so that all players would prefer coordinating on the same outcome which is most beneficial to everybody. On the other end of the spectrum are constant-sum games, in which players' interests are opposed: allocating a reward to one player is equivalent to taking it away from the other player(s). If a game combines the two features, so that players have conflicting incentives to cooperate, as they do in coordination games, and to compete, as in constant-sum games, it is deemed to be a game with mixed motives.

My findings can be summarised as follows. On the one hand, the combined effects of belief updating and aggregation suggest that groups can indeed be expected to coordinate better than individuals in the coordination games investigated. Interestingly, belief updating can lead rational Bayesian agents in small groups to over-estimate the probability of coordinating successfully which then turns into a self-fulfilling prophecy. The intuition behind this finding is that while fellow group members can be representative of the other partipants in the experiment, groups are too small for fellow group members to be a representative sample. On the other hand, there is no consistent difference between group- and individual behaviour in zero-sum games and games with mixed motives, and effects based on aggregation and belief updating cannot explain observed differences and similarities in particular games. The findings suggest that in order to understand group behaviour in such environments, more research is needed on how the group decision setting affects social preferences.

3.2 The Data

This review examines experimental studies that compare the behaviour of three-player groups to that of individuals in strategic or interactive decision tasks published between 1998 and 2011. Two exclusion conditions apply. One, I do not consider the game known as the *prisoners'* *dilemma* or variations thereof. There exists a large literature in social psychology which examines group behaviour in this particular decision task and which has been reviewed relatively recently in Wildschut et al. (2003) and Wildschut and Insko (2007). Two, I consider only the first round of play in repeated games and include only those games in which a single round of play has a meaningful interpretation. On the basis of this condition, I exclude Gillet et al. (2009)'s study of group behaviour in a *common resource pool game*.

The criteria are met by experimental investigations of six different coordination games and nine mixed-motive and constant sum games. The coordination games include two parametrisations of *weakest-link* or *minimum games* and four versions of *average-opinion* or median games, including the *continental divide* or *separatrix game* (Feri et al., 2010). Mixed-motive and constant sum games, on the other hand, include the *ultimatum game* (Bornstein and Yaniv, 1998), two versions of the *centipede game* (Bornstein et al., 2004), the *power-to-take game* (Bosman et al., 2006), the *investment* or *trust game* (Kugler et al., 2007), the *gift-exchange game* (Kocher and Sutter, 2007), the *beauty-contest* or *guessing game* (Kocher and Sutter, 2005) and the *dictator game* (Luhan et al., 2009). Each of these games constitutes a different decision task. Some are symmetric, so that all players in the game have the same set of possible actions. Some games are simultaneous, with all players making their decision at the same time, while others are sequential. In what follows, I will briefly explain the decision tasks involved in the different games.

Consider first the coordination games. In all six of these games, players simultaneously choose a number, and their pay-offs depend on the number they have chosen as well as a statistic of the numbers chosen by the other players. Both the possible actions, that is, the numbers that players can choose, and the pay-offs associated with them are the same for all players, so the games are symmetric. How pay-offs are determined depends on the specific game. In the *weakest-link* games, a player's pay-offs increase in the minimum number chosen by any of the players, but weakly decrease in his or her own number for a given minimum. The best response to a given strategy combination of the other players is to match the action of the player who has chosen the lowest number. In the *average-opinion* games, a player's pay-off depends on the median number chosen between all players, rather than the minimum. The further away a player's chosen number from the median number, the lower the payoff. The best response is then to match the action of the player whose chosen number is this median number. By definition, in all coordination games, players' interests are perfectly aligned so that all players would prefer coordinating on the same outcome which is most beneficial to everybody.

The mixed-motive and constant-sum games considered, on the other hand, are generally not symmetric nor simultaneous. Many of the games feature a mixed-motive decision for one player and a constant-sum decision for the other, so that rather than distinguishing between mixed-motive and constant-sum games, I consider the decisions made by the different players separately. In the well-known *ultimatum game*, for example, two players have to divide a fixed sum of money between them. The first player proposes a division which allocates a share of the sum to each of the two players, and the second player then decides to accept or reject this allocation. If it is accepted, each player receives the share proposed by the first player. If it is rejected, neither of the players receives anything. In this decision task, the decision faced by the first player is constant-sum, since the shares allocated to the players always have to add up to the initial sum. The decision made by the second player, on the other hand, is mixed-motive: the total amount allocated to the players can either be equal to the initial sum if the second player accepts the proposal, or equal to zero if it is rejected.

A number of games can be seen as variations on the game-theoretic structure of the ultimatum game. These include the *dictator game*, the *trust game*, the *gift-exchange game* and the *power-to-take game*. The structure of the dictator game, for example, is identical to that of the ultimatum game, with the exception that the second player does not have a decision to make and always has to accept the first player's proposal. In trust- and gift-exchange games, the first player receives a sum of money which increases if passed on to the second player. The second player then decides how much of the money, if any, to return to the first player. Here, the first player faces a decision with mixed motives, whereas the decision of the second player is constant-sum. Finally, in the power-to-take game, both players earn a sum of money, and the first player can decide to take part of the second player's money away from the latter. The second player then faces the decision whether to destroy any of the money beforehand, so that both players lose money.

In the *guessing game*, a large number of players simultaneously have to choose a number in a given interval. The winner is the player whose number is closest to a pre-determined fraction (a third, for example) of the average of the numbers chosen by all players in the game. The structure of this game resembles that of a coordination game, but players are competing with each other rather than coordinating. Last, in the *centipede game*, two players alternately get a chance to take the larger portion of a pile of money. As soon as a player 'takes' a portion, the game ends with that player getting the larger portion of the pile, while the other player gets the smaller portion.

Finally, although in some of the experiments games are played repeatedly, I am only concerned with behaviour in the first instance of the game. The goal of this study is to investigate a number of basic differences between individual and group decision settings. How learning processes differ between groups and individuals is an interesting but more complex question. The basic differences considered here may also affect more complex processes, however, so ideally, the findings of the present analysis could inform future work on group learning.

3.3 The Individual vs. Group Paradigm

My analyses are based on experimental results comparing the behaviour of individuals with that of unitary three-player groups, that is, with groups whose members jointly take the same action (Bornstein, 2003). Since the experimental method employed is largely the same across the studies which are the subject of the analyses in this paper, I will begin by briefly outlining this methodology.

At the beginning of the experiment, the rules of a particular decision task are explained to the participants. The participants then work on the decision task, either individually in the *Invidual* condition, or in groups of three players in the *Group* condition. The allocation of participants to these two conditions constitutes the main experimental manipulation. Participants in the *Individual* condition interact with other participants who act individually, whereas groups interact with other groups. Individuals in the *Group* condition are told that they have to arrive at a joint group decision but are free to make this decision by their own chosen method; decision rules are usually not prescribed. While communication is permitted between group members, who are often in each other's physical presence and allowed to interact with each other freely, there is no communication and maximal anonymity between different groups in the group setting, and between different individuals in the individual setting. Participants' choices in the experiment are monetarily incentivised, where group payoffs are three times individual payoffs, to be divided evenly among the group members to ensure that monetary incentives are the same across conditions. Finally, results are analysed by comparing the average behaviour of individuals in the individual setting with the average behaviour of groups in the group setting, in terms of the choice variable of the particular decision task.

$$V_i = v_0\left(\pi_0, \{\pi_0, \pi_1\}, b_0\right) \tag{3.1}$$

$$V_G = \mathbf{f}\left(v_0\left(\pi_0, \{\pi_0, ..., \pi_5\}, b_0(v_1, v_2)\right), v_1(...), v_2(...)\right) \tag{3.2}$$

Consider a strategic decision involving two individuals or two three-player groups, then. Equations 3.1 and 3.2 formalise three basic dimensions on which individual- and group conditions differ in this experimental paradigm. Individual and group choice, V_i and V_G, are determined by the preferred actions of the individuals v_i, which are combined by an aggregation function $\mathbf{f}$ in the group condition. Individuals' preferred actions v_i depend on preferences over the possible outcomes π_i for themselves and all other participants, as well as their beliefs b_i.

The equations demonstrate that the experimental manipulation of substituting groups in the place of individuals as decision makers in a game theoretic context is rather complex and leads to a number of differences between the conditions. First, they show that individual i's preferred action, V_i, depends on i's possible payoffs, π_0, as well as the vector of possible allocations to all players in the game. In the individual setting, this vector is comprised of only two individuals' possible payoffs $\{\pi_0, \pi_1\}$ and in the group setting of six individuals' possible payoffs $\{\pi_0, ..., \pi_5\}$, which marks the first difference between the two conditions. The second difference between individual- and group choice is that while in the individual setting, the final choice is simply the individual's preferred action, $V_i = v_0$, in the group setting, the group choice is the result of combining the preferred actions of the three group members, $V_G = \mathbf{f}(v_0, v_1, v_2)$. Finally, there is a difference in how beliefs are formed: while participants must rely on introspection to form beliefs b_0 in the individual setting, they can observe their fellow group members in the group setting and update their beliefs based on their behaviour, so that $b_0(v_1, v_2)$. The analyses in this paper assess in how far the differences between the decision settings can jointly explain observed differences between group and individual behaviour.

3.4 Analyses

In the above discussion of the "individual vs. group paradigm", I have identified three dimensions on which the individual and the group decision settings differ. They include how beliefs are formed, how many people are involved in the choice situation and how this affects decision makers' social preferences, and how group decisions are the result of aggregating the group members' preferred actions. I will now analyse these effects more formally in the context of the strategic decision tasks which the various games represent.

Table 3.1: Effects that affect group behaviour by game and decision

Game	Decision	Type†	B†	A†	S†
Weakest Link	Number chosen	Coor.	√	√	
Average Opinion	Number chosen	Coor.	√	√	
Continental Divide	Number chosen	Coor.	√	√	
Gift Exchange	Amount returned	C. Sum		√	√
Trust Game	Amount returned	C. Sum		√	√
Dictator Game	Other-allocation	C. Sum		√	√
C. Sum Centipede	Withdrawal round	C. Sum	√	√	√
Guessing Game	Number chosen	C. Sum	√	√	√
Gift Exchange	Amount sent	Mixed	√	√	√
Trust Game	Amount sent	Mixed	√	√	√
Ultimatum Game	Amount offered	Mixed	√	√	√
Inc. Sum Centipede	Withdrawal round	Mixed	√	√	√
Power-to-Take	Take rate	Mixed	√	√	√
Power-to-Take	Destruction rate	Mixed		√	√

† *B* stands for belief-updating effects, *A* for aggregation effects, and *S* for the effects of social preference. *Type* classifies the different decisions in a game as coordination, constant sum or mixed-motive.

Not every effect can be analysed in every game, however. In the ultimatum game, for example, the first player's offers are partly determined by his or her beliefs about what offer levels the other player would be willing to accept. Since the only behavioural measure is the actual offer level observed, these beliefs cannot be measured separately from an 'altruistic' desire to give to, or share with the second player. Table 3.1

shows which decisions in which games are subject to which of the three effects. I first examine aggregation effects, which can be quantified in all of the games considered. I then turn to the effect of belief updating, where I will focus on coordination games in which beliefs are not confounded with social preferences. Finally, I discuss the impact that social preferences are likely to have in mixed-motive and constant-sum games, paying special attention to those decisions which do not depend on beliefs about the other players' behaviour.

3.4.1 Aggregation effects

The first step in my analysis examines aggregation effects in group behaviour. Equations 3.1 and 3.2 show how they depend on the aggregation function **f**, which describes how the actions v which group members would have chosen on their own translate to a group decision. Furthermore, aggregation effects can be observed in all of the games described in Section 3.2 (cf. Table 3.1). I will now examine whether aggregation leads to systematic effects, assuming that groups use a version of the "majority-rule" aggregation function. In order to make predictions about group behaviour, I use data on behaviour in the *individual* choice setting. The analyses show that aggregation effects are consistent with groups coordinating better with one another in coordination games than individuals, and that aggregation effects cannot explain behavioural differences between groups and individuals in mixed-motive and constant-sum games.

The nature and the size of a possible aggregation effect when comparing individual decisions with group decisions depends, of course, on the functional form of the aggregation function **f**. Consider, for example, a managerial group whose three members disagree on how much of a budget to spend on a particular project: two managers want to allocate 50 per cent of the total budget to the project, while the third thinks the project will fail and does not want to allocate any funds to it. If they were to decide by simply taking the average of their three opinions, they would allocate one third or 33.3 per cent of the budget. If, on the other hand, they were to decide by simple majority, the project would receive 50 per cent of the budget.

Majority rules

Social scientists have studied different aggregation functions and their properties since the 18^{th} century (see, for example, Condorcet, 1785; Arrow, 1963; Sen, 1977; Hastie and Kameda, 2005), but for the purpose of my analyses, I will restrict myself to a single aggregation function. In recent years, social psychologists have investigated the decision rules that groups use when left to their own devices, and have found that a large majority of groups use a version of the "majority-wins" rule (Davis, 1973; Laughlin and Ellis, 1986; Crott et al., 1991; Kameda et al., 2003). The only paper that investigates this question in the context of the Individual-vs.-Group paradigm echoes this result (Bosman et al., 2006), so in what follows I assume that the groups decide by "majority-wins" rules.

Aggregating individual behaviour

In order to quantify aggregation effects, I first compile information on individual behaviour in the individual choice setting from the sources given in Section 3.2. The required information includes choices and their frequencies: in Bornstein and Yaniv's (1998) experiment on group and individual behaviour in the ultimatum game, for example, 10 per cent of participants in the individual choice setting chose to offer thirty per cent of the pie, 20 per cent made offers of forty per cent of the pie, 65 per cent offered half the pie, and 5 per cent of participants offered sixty per cent. For three of the fourteen games, including the dictator-, trust-, and guessing games, the source in Section 3.2 does not include the information at the level of detail required. For these games, I adopt individual choices and their frequencies from previous studies: Forsythe et al. (1994) for the dictator game, Berg et al. (1995) for the trust game, and Duffy and Nagel (1997) for the guessing game.

Given these 'distributions' of individual behaviour, I construct all possible hypothetical groups of three players and apply a majority-wins rule to them according to which the group always adopts the median of the choices proposed by its members. A group deciding on an offer in an ultimatum game whose three members propose to offer 20, 30, and 50 per cent of the pie, will make a group offer of 30 per cent, for example. I then calculate an expected value μ_0 of group behaviour by multiplying the probability with which a particular group composition occurs with its group choice and summing over all possible group constellations. This expected value can be interpreted as a prediction concerning group behaviour in the absence of any effects other than aggregation. As such, this prediction excludes belief-updating and the effects of the

Table 3.2: Observed group behaviour and how it differs from aggregated individual behaviour

		Experimental Data			T-test: Mean = μ_0	
Game	Decision	Mean	StDev	N *	μ_0	p
Weakest Link: Base	Number chosen	6.51	1.01	45	6.35	0.31
Weakest Link: Risk	Number chosen	6.37	1.54	30	5.69	0.02
Average Opinion: Base	Number chosen	6.17	0.95	30	5.79	0.04
Average Opinion: Risk	Number chosen	4.40	1.04	30	4.20	0.29
Average Opinion: Pay	Number chosen	6.43	0.97	30	5.39	0.00
Continental Divide	Number chosen	11.0	2.91	30	7.69	0.00
Gift Exchange (1)†	Amount returned	0.20	0.12	24	0.21	0.70
Gift Exchange (2)†	Amount returned	0.43	0.30	16	0.25	0.03
Trust Game	Amount returned	20.7	20.4	52	24.0	0.26
Dictator Game	Other-allocation	0.11	0.11	30	0.19	0.00
C. Sum Centipede	Withdrawal round	2.00	0.75	18	2.55	0.01
Guessing Game	Number chosen	30.1	16.7	35	31.7	0.77
Gift Exchange (1)†	Amount sent	33.0	16.0	24	48.3	0.00
Gift Exchange (2)†	Amount sent	48.1	23.2	16	48.3	0.97
Trust Game	Amount sent	48.7	42.0	52	74.5	0.00
Ultimatum Game	Amount offered	0.36	0.12	20	0.48	0.00
Inc. Sum Centipede	Withdrawal round	4.44	0.96	18	5.16	0.01
Power-to-Take	Take rate (%)	60.0	13.8	12	61.5	0.71
Power-to-Take	Destruction rate (%)	20.8	39.6	12	9.22	0.33

† Data on behaviour in the gift-exchange game exists for two different conditions in which group members interact with each other through computers or in person (Kocher and Sutter, 2007). Results differ between the conditions, so both are included here with (1) referring to the computerised condition.

* N denotes the number of groups or observations.

group setting on social preferences, but also more complex group processes that may occur in the group setting such as persuasion effects and social comparison processes. Finally, I test the null hypothesis that the group behaviour observed in the experiments does not differ from this predicted behaviour by means of a T-test. The results of this analysis are shown in Table 3.2.

For eight of the nineteen decision tasks, the null hypothesis cannot be rejected at the five per cent level of statistical significance. If one adopts a more stringent significance level of one per cent, the null hypothesis cannot be rejected in eleven of the nineteen decisions. In all but one of these decision tasks (the *Weakest Link: Base* game in Feri et al. (2010)), the source papers cited in Section 3.2 do not find a significant difference between group behaviour and *unaggregated* individual behaviour, either. Overall, the extent to which group behaviour differs from individual behaviour is therefore limited.

Discussion

The absence of a consistent difference between group and individual behaviour across the variety of decision tasks considered is the main conclusion that can be drawn from Table 3.2. Nonetheless, it must be qualified both in terms of the decision tasks and in terms of the direction of the observed differences between group and and individual behaviour when these are present.

First, consider only the six coordination games in the first six rows of the table. In these games, the behavioural difference is more consistent than in the constant-sum and mixed-motive games. At a significance level of five per cent, groups choose higher numbers than predicted by aggregating individual behaviour in four of six games, and in a fifth game, the aggregation effect explains the difference between group and individual behaviour. Furthermore, the effect in all six games is in the same direction: the numbers groups choose are on average higher than those predicted by aggregating individual behaviour, and as a result, they are able to coordinate more often on the efficient outcome (not shown, see Feri et al., 2010). In the next section, I argue that this behavioural difference can be partly explained by how belief-updating affects group decisions.

Second, when a difference exists between groups and individuals in constant-sum and mixed-motive decisions, in all but one case, groups behave more selfishly than individuals. Only in the non-computerised condition of the gift-exchange game in Kocher and Sutter (2007) do groups behave less selfishly than individuals, returning more to the other player in the game than participants in the individual choice setting. Table 3.2 therefore shows that there is a tendency for groups to behave selfishly, but that this effect cannot be found consistently across decision tasks. I return to this issue when discussing the effects of social preferences on group decisions below, but cannot provide a conclusive account of this inconsistency.

3.4.2 Belief updating in coordination games

Coordination games represent strategic decisions in which the interests of all parties who are involved in the decision are perfectly aligned. All parties therefore have an incentive to coordinate on a particular outcome, but may not always be able to do so because they do not know how the other parties will behave. Furthermore, if the strategic decision is symmetric, like in the coordination games considered here, all parties receive the same pay-off if they coordinate successfully. Since everybody's interests are aligned, behaviour will not depend on the decision makers' preferences over the pay-offs of others as captured by social preferences such as inequality aversion (Loewenstein et al., 1989; Fehr and Schmidt, 1999) or other social preferences that have been studied in the literature (Loewenstein et al., 1989; Charness and Rabin, 2002; Fehr and Fischbacher, 2002). Now re-consider Equations 3.1 and 3.2 in the environment of such a coordination game. If social preferences do not influence behaviour, the equations can be simplified.

$$
\begin{aligned}
V_i &= v_0\left(\pi_0, b_0\right) \\
V_G &= \mathbf{f}\left(v_0\left(\pi_0, b_0(v_1, v_2)\right), v_1(\ldots), v_2(\ldots)\right)
\end{aligned}
$$

In the simplified equations, one dimension on which the individual- and the group choice settings differ has disappeared. If social preferences do not affect behaviour in either setting, they cannot influence behaviour in the two choice settings differently, either. Instead, the simplified equations suggest that behavioural differences between individuals and groups could be due to differences in belief formation and aggregation effects.

This implies that behaviour in the *individual* decision setting will be determined almost exclusively by participants' beliefs about the other participants' behaviour. Since these decisions are made in isolation, beliefs about others' behaviour must be based on introspection. In the group decision setting, on the other hand, each participant can observe the behaviour of two other participants, his or her fellow group members. Since these belong to the same subject pool as the participants in the other groups that they are coordinating with, they are representative of these other participants. It is therefore sensible for participants to update their beliefs about other participants' behaviour based on the behaviour they observe in their own group.

I analyse the implications of this belief-updating process in a simple Bayesian model which will be formalised below. Participants know which action v they would prefer to take on their own. Furthermore, they have prior beliefs b about the probabilities with which a second, randomly chosen participant will prefer to take each of the possible courses of action. In the group decision setting, the members of a group announce their preferred action v to one another in order to make a group decision. From a particular participant's point of view, the actions that his or her fellow group members announce they would take are two realisations from an unknown distribution over the v. (S)he can then update his or her beliefs b, taking into account the preferred actions of the group mates. My analysis shows that if priors b are unbiased, but are sufficiently weak for the updating process to have an effect, posteriors regarding the other participants' behaviour will differ from b and will hence not be unbiased anymore. Furthermore, in four of five games, the posterior distributions of the majority of participants assign a higher probability to others choosing the high-efficiency outcome than the rational prior, which could explain why groups may be able to coordinate better than individuals.

A Weakest-link game (from Feri et al., 2010)

Consider, for example, the *WL-Risk* game reported in Feri et al. (2010). Five players simultaneously announce a number between one and seven, and players' pay-offs in this coordination game are determined by the number they announce and the smallest of the five numbers. As shown in Table 3.3, players whose chosen number coincides with this minimum number obtain a positive pay-off. This pay-off increases with the minimum number: the larger the number on which players coordinate, the higher the pay-offs for those players who have chosen this number. Players whose chosen number differs from the minimum number receive a pay-off of zero, on the other hand.

Table 3.3: Pay-offs in the WL-Risk game from Feri et al. (2010)

	Smallest # chosen by other players						
Own #	1	2	3	4	5	6	7
7	130	0	0	0	0	0	0
6		120	0	0	0	0	0
5			110	0	0	0	0
4				100	0	0	0
3					90	0	0
2						80	0
1							70

It is in all players' interest to coordinate on the first row, with everybody choosing the number seven, because it maximises everybody's pay-off. As soon as one of the players chooses a smaller number, however, all players that choose the number seven receive a pay-off of zero. Clearly, players' choices depend crucially on their beliefs about the behaviour of other players. In particular, players will be more likely to choose the number seven and coordinate successfully on the high-efficiency outcome in the first row when they believe that other players will also choose the number seven. Belief-updating will then make groups more likely to coordinate successfully if group members' posteriors assign higher probabilities to other players choosing the number seven than their priors.

I will now formalise the Bayesian model of belief-updating outlined above in the context of the *WL-Risk* game. First, assume that each participant prefers a number $v_i \in \{1, 2, 3, 4, 5, 6, 7\}$ with probability $p_j = \mathbb{P}\left(v_i = j\right)$. Participants' preferred numbers v_i can then be thought of as independent realisations from a probability distribution f which follows the Categorical distribution $f \sim Cat(\mathbf{p})$, where $\mathbf{p}$ denotes the vector $\{p_1, p_2, ..., p_7\}$. In analogy to my analysis of aggregation effects above, I construct the probability vector $\mathbf{p}$ from the data on individual behaviour in the *individual decision setting* of the game[1]. Forty-seven per cent of participants in the individual decision setting chose the number 7, for example, and only seven per cent of participants the number 1. The complete vector is given by $\mathbf{p} = \{0.07, 0.03, 0.1, 0.1, 0.13, 0.1, 0.47\}$. Since the participants in the individual- and group decision settings of the game were recruited from the same subject pool, this probability vector should describe the preferences of the participants in the group decision setting rather well.

In a Bayesian framework, participants' beliefs b about other participants' behaviour can be thought of as priors for the distribution f. These priors can be modelled conveniently as following a Dirichlet dis-

[1]Many thanks to Matthias Sutter for sharing the data from the Feri et al. (2010) experiments with me.

tribution $b \sim Dir(\alpha)$, which is conjugate to the Categorical distribution. I assume that priors are unbiased and reflect the true distribution f by choosing the parameters α such that $\mathbb{E}\left[b_j\right] = p_j \forall j$. Before interacting with one another in the groups, participants therefore believe that a second participant chosen at random will choose the number seven with a probability of on average $\mathbb{E}[b_7] = p_7 = 0.47$, for example. While I assume that priors are unbiased, I do not assume these prior beliefs to be strong, however. This is equivalent to saying that participants are able to predict others' behaviour in the game correctly on average, but will revise their beliefs substantially upon observing the behaviour of a sample of participants, even if this sample is small.

How do participants revise these beliefs upon observing what number their fellow group members would prefer to choose? From the point of view of a given participant in the experiment, the two numbers which reflect the preferred actions of the other two group members (v_1, v_2) can be thought of as two independent realisations from the distribution f. On the basis of this new information they have acquired about f, participants can then revise their beliefs. The resulting posterior belief $b \mid (v_1, v_2)$ will change with respect to the intial belief b, assigning higher probabilities to those numbers preferred by the other group members, and lower probabilities to the numbers that are not chosen by either one of the other group members. A participant in a group in which both other group members prefer to choose the number seven, for example, would, as a result, believe that the probability of other participants choosing the number seven is higher than initially assumed. In particular, the posterior vector of beliefs $b \mid (v_1 = 7, v_2 = 7)$ is given by $\mathbb{E}[b \mid (v_1 = 7, v_2 = 7)] = (0.02, 0.01, 0.03, 0.03, 0.04, 0.03, 0.82)$.

Note that the posterior vector $b \,|(v_1 = 7, v_2 = 7)$ is not unbiased anymore: the probabilities that a participant assigns to other participants choosing a particular number do not reflect those of the true distribution anymore. Furthermore, participants whose fellow group members would both prefer to choose the number 7 are *more optimistic* as a result of the group interaction. The probability that their posterior beliefs assign to others choosing the number 7 $\mathbb{E}\,[b_7|(v_1 = 7, v_2 = 7)] = p_7 = 0.82$ associated with the high-efficiency outcome has increased with respect to the prior $\mathbb{E}\,[b_7] = p_7 = 0.47$. As a consequence, such participants will be more likely to choose the number 7 themselves.

How likely is any participant to be in a group with two others who prefer to choose the number seven? Based on the distribution f, the probability $\mathbb{P}(v_1 = 7, v_2 = 7)$ can be calculated as $\mathbb{P}(v_1 = 7, v_2 = 7) = \mathbb{P}(v_1 = 7) * \mathbb{P}(v_2 = 7) = 0.47^2 = 0.22$. Twenty-two per cent of participants in the group decision setting can therefore be expected to find themselves in a group in which both their fellow group members prefer to choose the number seven. Finally, consider all possible group constellations. For each constellation, I calculate the probability with which it occurs $\mathbb{P}(v_1 = v, v_2 = u) = \mathbb{P}(v_1 = v) * \mathbb{P}(v_2 = u)$. I then calculate the posterior belief of a hypothetical participant exposed to each of the possible group constellations $b \,|(v_1 = v, v_2 = u)$, and compare the probability that the posterior beliefs assign to others choosing the number seven associated with the high-efficiency outcome, $\mathbb{E}\,[b_7|(v_1 = v, v_2 = u)]$ to that of the prior $\mathbb{E}\,[b_7]$. If the posterior probability is larger than the prior probability, I consider the hypothetical participant *more optimistic* regarding the probability of coordination on the high-efficiency outcome as a result of observing the other group members and updating his or her beliefs accordingly. If, on the other hand, the posterior belief assigns a lower probability to others choosing the number seven, the group decision setting has made the participant less optimistic.

Finally, I compare the expected proportion of hypothetical participants who become more optimistic as a result of the belief-updating process in the group decision setting[2] with the proportion of those who become less optimistic. This analysis shows that in the *WL-Risk* game reported in Feri et al. (2010), if the participants in the individual decision setting are representative of those in the group decision setting, and if participants in the group decision setting update their beliefs as proposed here, seventy-two per cent of participants can be expected to be more optimistic as a result of this rational belief-updating process. Only thirty-eight per cent of participants can be expected to assign a lower probability to others choosing the number seven after observing their fellow group members' behaviour and updating their beliefs. In the *WL-Risk* game reported in Feri et al. (2010), rational belief updating processes therefore provide a new explanation for why groups may be able to coordinate better than individuals.

Other coordination games

The experiments reported in Feri et al. (2010) compare individual and group behaviour in five additional coordination games. Four of these games have pay-off structures which resemble that of the *WL-Risk* game: players have to choose a number from one to seven, and their incentives are aligned so that it is in their interest to coordinate on choos-

[2]Calculated as $\mathbb{P}_{opt} = \sum \mathbb{P}(v_1 = v, v_2 = u) * \mathbb{I}(\mathbb{E}[b_7 | (v_1 = v, v_2 = u)] > \mathbb{E}[b_7])$.

ing the same number in order to maximise their pay-offs[3]. For each of these games, I conduct an analysis of belief-updating analogous to the one above. The results, shown in Table 3.4, show that for four out of five games, belief-updating processes will make participants more optimistic regarding the possibility of coordinating on the high-efficiency outcome in the group decision setting, with the only exception being the *WL-Base* game. Note that this corresponds to the only coordination game in which differences between group and individual behaviour can be explained by a pure aggregation effect.

Table 3.4: Belief-updating in the coordination games in Feri et al. (2010)

	Game				
	WL-Base	WL-Risk	AO-Base	AO-Risk	AO-Pay
$\mathbb{P}_{opt}$	0.37	0.72	0.72	0.59	0.60

$\mathbb{P}_{opt}$ denotes the probability with which participants in the group decision setting will assign a higher posterior probability to other players coordinating on the high-efficiency outcome as a result of the belief-revision process.

Discussion

My analysis suggests that group-based Bayesian belief-revision processes can alter decision makers' beliefs regarding the behaviour of others. In four out of five of the coordination games considered, I find that observing fellow group members' behaviour makes decision makers more optimistic with respect to the possibility of coordinating successfully.

[3]The seventh game, *Continental Divide*, has a larger strategy space and has a component of problem-solving which is not present in the other games. Since these additional considerations make the link between players' beliefs about other players' behaviour and the actions they take weaker, I exclude the game from the analysis.

Furthermore, I obtain these findings assuming that priors are unbiased, providing a rational explanation for why groups may be able to coordinate better than individuals.

This result rests on two major assumptions. One, the group size is limited, so that decision makers update their beliefs on the basis of a small sample. If groups were large, and decision makers could update their beliefs taking into account the behaviour they observe in all their fellow group members, their posteriors would not differ substantially from their unbiased priors. Two, priors are sufficiently weak for the updating process to have an effect on the posterior, even though the small sample of behaviour that decision makers observe in their fellow group members provides relatively little information. If prior beliefs were so strong that a small sample of two new observations would not have an effect on them, the belief-revision effect would disappear[4].

In the context of the experimental work analysed here, both of these assumptions are satisfied. On the one hand, the Individual-vs.-Group paradigm examines the behaviour of small groups which consist of exactly three group members. On the other, playing coordination games in an experimental laboratory probably constitutes a novel decision task for experimental participants, who will usually have little experience in playing these exact games. Their prior beliefs about other participants' behaviour can be expected to be relatively weak.

[4]The size and the direction of the belief-updating effect also depend on how individuals behave on their own as described by the distribution f. For belief-updating to lead to greater optimism, rather than pessimism, regarding the possibility of successful coordination, f needs to meet additional distributional restrictions. This explains the effect toward greater pessimism in the *WL-Base* game shown in Table 3.4.

3.4.3 Social preferences in group decisions

The third and final difference between the group- and the individual decision setting is the number of people who are affected by the outcome of a decision, and the relations in which they stand to one another. Decision makers who care about the outcomes of others exhibit *social preferences*. For these decision makers, a decision task involving a single other person like in the individual decision setting differs from one involving five others, as in the group decision setting. My analysis shows that while there is a general tendency for groups to behave more selfishly than individuals in constant-sum and mixed-motive decision tasks, the evidence from pure allocation decisions which depend only on social preferences is mixed.

Pure Allocation Decisions

I argued above that in many strategic decision tasks, behaviour is jointly determined by social preferences and beliefs about other players' behaviour and that these two cannot be distinguished from one another. Players may make generous offers in the ultimatum game, for example, because they have an altruistic desire to give, or their generous offers may reflect their beliefs that lower offers would be rejected by the other player. The decision tasks considered here include five pure allocation decisions, however, in which uncertainties about other players' behaviour, if they exist, are resolved before the decision has to be made. For these decisions, Equations 3.1 and 3.2 can again be simplified.

The results reported in Table 3.2 in such decision tasks can be interpreted as (crude) tests of the hypothesis that the group decision setting does not affect social preferences. They include the amounts returned in the gift-exchange and trust games, the amount allocated to others in the dictator game, and the destruction rate in the power-to-take game (cf. Table 3.1). If individuals making decisions as part of a group were more selfish than when making decisions on their own, group behaviour should be more selfish than aggregated individual behaviour in all of these decision tasks. Table 3.2 shows that this is not the case: it is only in the dictator game that groups allocate less to the other players than individuals. In the computerised condition of the gift-exchange game, the trust game, and the power-to-take game, there is no difference between group and (aggregated) individual behaviour. Finally, in the non-computerised condition of the gift-exchange game, groups even return a larger amount to the other player than individuals.

$$V_i = v_0\left(\pi_0, \{\pi_0, \pi_1\}\right)$$
$$V_G = \mathbf{f}\left(v_0\left(\pi_0, \{\pi_0, ..., \pi_5\}\right), v_1(...), v_2(...)\right)$$

Allocation decisions with uncertainty

Nonetheless, when beliefs and social preferences interact in strategic decision tasks, the tendency for groups to behave more selfishly than individuals reappears. Although not a consistent effect across all eight instances of such decision tasks, groups behave significantly more selfishly than individuals in five of them. This finding confirms the 'discontinuity effect' (McCallum et al., 1985; Insko et al., 1990, 1994) which holds that intergroup interactions are often more competitive and less cooperative than interindividual interactions. Unlike the experiments examined here, the discontinuity effect has been replicated numerous times using variations of the well-known prisoners' dilemma game (Wildschut et al., 2003; Wildschut and Insko, 2007). In a recent review of the different explanations that have been proposed to explain the effect, Wildschut and Insko (2007) suggest that the evidence is strongest in favour of accounts based on norms favouring the in-group (Tajfel and Turner, 1979), on social support by the group for acting selfishly (Insko et al., 1990) and on the anonymity provided by the group (Schopler et al., 1995). Wildschut and Insko (2007) refer to these theoretical explanations as the *fear and greed perspective* on the discontinuity effect. That groups are not unequivocally more selfish in pure allocation decisions in the experiments reviewed here can be seen as favouring fear-based explanations over greed-based ones. It is also consistent with recent experiments which find that in-group altruism is more important in explaining individual behaviour in group settings than aggression toward the out-group (Halevy et al., 2008; Chen and Li, 2009).

There also exist different formalisations of social preferences which have been developed by behavioural economists. Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002), among others, provide models in the tradition of utility theory designed to describe the behaviour of decision makers with preferences about allocations to others in addition to allocations to themselves. They explore concepts such as difference aversion, which makes decision makers favour equal allocations over unequal ones, or let decision makers be concerned with maximising the social welfare of others. These formal models of social preferences can be used to make behavioural predictions in decision tasks which involve the decision maker and one other person, but it is not clear how they would generalise to environments with more than two individuals like the group decision settings analysed here[5]. In a two-person setting, difference aversion can be readily defined with respect to the difference in the allocations of the two individuals, for example. It is unclear how the concept translates to a three-person environment, however: would decision makers be averse to differences between the largest and the smallest allocation, or between the largest and the next largest allocation? Or would they only be averse to differences between their own allocation and others' allocations? In order to make meaningful generalisations to choice settings with multiple individuals which could help understand group behaviour, the psychological processes underlying concepts such as difference aversion or social-welfare-maximisation need to be better understood.

Finally, recent results show that other-regarding behaviour and social preferences are subject to context effects. Choshen-Hillel and Yaniv (2011) identify decision makers' agency, the degree to which they influence, control or create the outcomes of a decision as a key determinant of pro-social behaviour. In a series of experiments, they show that non-agentic decision makers who cannot influence the allocations to others are concerned with avoiding inequalities, whereas agentic decision makers with a degree of control over the allocation are more likely to maximise the social welfare of others. How making decisions jointly as part of a group affects decision makers agency is still unclear, but it is another factor which could help our understanding of group behaviour in the games reviewed here.

[5]Charness and Rabin (2002) include a mathematical generalisation to n-person environments in an Appendix, but voice their concerns regarding the psychological validity of this generalisation.

3.5 Conclusion

Many important strategic decisions in firms, agencies and other organisations are made by comittees or other small groups, and not by individuals. How groups behave in different game-theoretic settings is therefore of interest to social scientists, and in this essay, I have re-assessed previous experimental findings regarding group behaviour. I conduct my analyses in a simple theoretical framework of group behaviour which abstracts from the sophisticated psychological processes which decision makers experience when making decisions as part of a group. Restricting my framework to the game-theoretic context, to players' beliefs and preferences, and to aggregation effects, I investigate in how far these can jointly explain group behaviour. My objective is to partly formalise the group-decision-making process and to identify decision tasks in which knowledge of sophisticated psychological processes is not required in order to explain group behaviour. Conversely, and perhaps more interestingly, I also highlight decision tasks in which an effort should be made to understand these psychological processes.

One of the contributions of the present investigation lies in applying the theoretical insights from the literatures on social preferences (Loewenstein et al., 1989; Charness and Rabin, 2002; Choshen-Hillel and Yaniv, 2011) and aggregating opinions and preferences (Crott et al., 1991; Gigone and Hastie, 1997; Kameda et al., 2003) to experimental results concerning group behaviour in which they are likely to have important effects but had not yet been considered. In addition, it introduces a new theoretical aspect of group decision making in its analysis of belief-updating effects in groups. Decision makers who form part of a group may update their beliefs about the behaviour of other people from the behaviour they observe in their fellow group members. In groups which are limited in size, other group members will be too small a sample to be representative of the population they belong to. If prior beliefs are unbiased but sufficiently weak, the resulting posteriors can be biased as a consequence of the belief-updating process.

The experimental data reviewed here show that group behaviour in coordination games is largely consistent with the predictions of the framework. Aggregation effects and effects from belief-updating both suggest that groups should be able to coordinate better with each other than individuals. This adds to the observation that groups learn better in coordination games than individuals (Feri et al., 2010) and provides an explanation as to why groups have an advantage even when the game is played only once, or in the first round of repeated play. The learning-based account of group behaviour in coordination games should be re-evaluated in the light of this finding given the importance of the starting point in learning processes in strategic environments.

In mixed-motive and constant-sum games, on the other hand, the framework proposed here falls short of explaining group behaviour. The analysis suggests that in such environments, the differences between group behaviour and individual behaviour are less consistent, however. I examine a number of pure allocation decisions which are not influenced by decision makers' beliefs about others' behaviour, and in the five instances of such decisions considered here, there is no consistent difference between how groups behave and the behaviour of individuals. Group behaviour in pure allocation decisions highlights our limited knowledge about social preferences and other-regarding behaviour in decision settings which involve more than two people.

Yet, in the presence of uncertainties about others' behaviour, the experimental results reviewed here confirm what is known as the "discontinuity effect" between inter-group and inter-individual behaviour (Wildschut and Insko, 2007). Groups behave more selfishly, more competitively and less cooperatively than individuals in a majority of those decision tasks considered which depend both on beliefs and on social preferences. Together, the observations that groups are not unequivocally more selfish than individuals in pure allocation decisions but that they behave more competitively than the latter in the presence of uncertainty could lead to new insights about the discontinuity effect. For the prominent "fear and greed" perspective on the discontinuity effect, they emphasize the importance of fear, since groups do not behave greedily in pure allocation decisions.

The present analysis has a number of methodological implications. One, the analysis shows why experimental studies of group behaviour in strategic environments often exhibit a surprising degree of complexity. These studies combine a complex experimental manipulation (individuals vs. groups as decision makers, see Section 3.3) with decision tasks where outcome measures often confound more than one motivational factor (beliefs vs. preferences, see Section 3.2). This suggests that it may be worthwhile to conduct more studies of group behaviour in decision tasks which separate beliefs about how others will behave from 'pure' social preferences as determinants of the decision. Two, more research is needed on how making decisions in environments with multiple agents affects social preferences. This includes group decision making, but is not restricted to it. Take the dictator game, for example. What if dictators had to share their money with three other people instead of just one? An answer to this question probably requires an entire research programme, but would lay a foundation upon which the more complicated question of group behaviour in the dictator game could be investigated. Finally, three, it would be useful for future experimental work employing the Individual-vs.-Group paradigm to compare observed group behaviour to aggregated individual behaviour as well as unaggregated individual behaviour. The median-voter implementation of the "majority wins"-rule proposed here is defined for any dependent variable which can be ordered, is calculated easily, and the comparison allows researchers to discard aggregation as a possible explanation of any observed effects.

This essay also has implications for how to think about groups as strategic decision makers in organisations. The framework proposed in this paper shows how belief-updating in small groups may lead decision makers to believe that the group is overly representative of the population its members belong to, and this effect is not restricted to small groups in an experimental laboratory. Task forces or other teams formed within an organisation often count with ten or less group members. At the same time, such teams often have to make decisions which depend on the behaviour of other people in- or outside of the organisation. A marketing team may have to decide on the nature of an advertisement campaign, or an R&D team of engineers on the features to include in a new gadget. Both of these teams make good decisions if they are able to correctly predict the behaviour of their future clients. The present analysis shows that their predictions may suffer if team members think of each other as being representative of these future clients, because the teams are too small to be representative. On the other hand, experimental results regarding the competitiveness and selfishness of intergroup interactions should be taken with a grain of salt when generalising to teams as decision makers in organisations. Many decisions that have to be made in organisations cannot be classified as pure allocation decisions, or as depending as directly on uncertainties about others' behaviour as an ultimatum game played in a laboratory, but this difference determines group behaviour in the experimental data reviewed here. Groups that make decisions in organisations may also be of a different size than the group affected by the outcome of the decision, and it is not known how differences in group size affect social preferences or behaviour. In conclusion, this review suggests that more research is needed on group behaviour and its determinants in strategic environments to make accurate predictions about group behaviour in more complicated strategic decisions like the ones they may face in an organisation.

Bibliography

Armstrong, J. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers.

Arrow, K. (1963). *Social choice and individual values*. Yale University Press.

Becker, G., DeGroot, M., and Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, 8(1):41–55.

Bell, D. (1982). Regret in decision making under uncertainty. *Operations Research*, 30(5):961–981.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Block, H. and Marschak, J. (1960). Random orderings and stochastic theories of responses. *Contributions to Probability and statistics: Essays in Honor of Harold Hotelling*, pages 97–132.

Bolton, G. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review*, 7(2):129–145.

Bornstein, G., Kugler, T., and Ziegelmeyer, A. (2004). Individual and group decisions in the centipede game: "are groups more rational" players? *Journal of Experimental Social Psychology*, 40(5):599–605.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: Are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Bosman, R., Hennig-Schmidt, H., and van Winden, F. (2006). Exploring group decision making in a power-to-take experiment. *Experimental Economics*, 9(1):35–51.

Brunswik, E. (1956). *Perception and the Representative Design of Psychological Experiments*. University of California Press.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117(3):817–869.

Chen, Y. and Li, S. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–457.

Choshen-Hillel, S. and Yaniv, I. (2011). Agency and the construction of social preference: Between inequality aversion and prosocial behavior. *Journal of Personality and Social Psychology*, 101(6):1253–1261.

Colman, A. (1982). *Game theory and experimental games: The study of strategic interaction*. Pergamon Press Oxford.

Condorcet, M. (1785). *Essay sur l'application de l'analyse de la probabilité des decisions: Redues et pluralité des voix*. l'Imprimerie royale.

Crano, W. and Prislin, R. (2006). Attitudes and persuasion. *Annual Review of Psychology*, 57:345–374.

Crott, H., Szilvas, K., and Zuber, J. (1991). Group decision, choice shift and polarization in consulting, political and local political scenarios: An experimental investigation and theoretical analysis. *Organizational Behavior and Human Decision Processes*, 49(1):22–41.

Davis, H., Hoch, S., and Ragsdale, E. (1986). An anchoring and adjustment model of spousal predictions. *Journal of Consumer Research*, 13(1):25–37.

Davis, J. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, 80(2):97–125.

Delquie, P. and Cillo, A. (2006). Disappointment without prior expectation: a unifying perspective on decision under risk. *Journal of Risk and Uncertainty*, 33(3):197–215.

DeMarzo, P., Vayanos, D., and Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, 112(4):951–978.

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological Review*, 114(1):177–187.

Denrell, J. and Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, 114(2):398–422.

Denrell, J. and Le Mens, G. (2011). Seeking positive experiences can produce illusory correlations. *Cognition*, 119(3):313–324.

Dhami, M., Hertwig, R., and Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6):959–988.

Duffy, J. and Nagel, R. (1997). On the robustness of behaviour in experimental beauty contest games. *The Economic Journal*, 107(445):1684–1700.

Einhorn, H. and Hogarth, R. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85(5):395–416.

Einhorn, H. and Hogarth, R. (1986). Decision making under ambiguity. *Journal of Business*, 59(4):225–250.

Fehr, E. and Fischbacher, U. (2002). Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, 112(478):1–33.

Fehr, E. and Schmidt, K. (1999). A theory of fairness, competition, and cooperation*. *Quarterly Journal of Economics*, 114(3):817–868.

Feri, F., Irlenbusch, B., and Sutter, M. (2010). Efficiency gains from team-based coordination: large-scale experimental evidence. *American Economic Review*, 100(4):1892–1912.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological Review*, 107(4):659–676.

Fiedler, K. and Juslin, P. (2006). *Information sampling and adaptive cognition*. Cambridge University Press.

Forsythe, R., Horowitz, J., Savin, N., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Galton, F. (1907). Vox populi. *Nature*, 75:450–451.

Gigone, D. and Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin*, 121(1):149–167.

Gilbert, D., Killingsworth, M., Eyre, R., and Wilson, T. (2009). The surprising power of neighborly advice. *Science*, 323(5921):1617–1619.

Gilbert, D., Pinel, E., Wilson, T., Blumberg, S., and Wheatley, T. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3):617–638.

Gilbert, D. and Wilson, T. (2000). Miswanting: Some problems in the forecasting of future affective states. In Forgas, J., editor, *Feeling and thinking: The role of affect in social cognition*, pages 178–197. Cambridge University Press.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Gillet, J., Schram, A., and Sonnemans, J. (2009). The tragedy of the commons revisited: The importance of group decision-making. *Journal of Public Economics*, 93(5-6):785–797.

Halevy, N., Bornstein, G., and Sagiv, L. (2008). "in-group love" and "out-group hate" as motives for individual participation in intergroup conflict. *Psychological Science*, 19(4):405–411.

Harrison, J. and March, J. (1984). Decision making and postdecision surprises. *Administrative Science Quarterly*, 29(1):26–42.

Hastie, R. and Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, 112(2):494–508.

Hertwig, R., Barron, G., Weber, E., and Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539.

Herzog, S. and Hertwig, R. (2009). The wisdom of many in one mind. *Psychological Science*, 20(2):231–237.

Hogarth, R. (2010). On subways and coconuts in foggy mine fields: an approach to studying future-choice decisions. In Michel-Kerjan, E. and Slovic, P., editors, *The irrational economist: making decisions in a dangerous world*. Public Affairs.

Hsee, C. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 67(3):247–257.

Hsee, C. (2000). Attribute evaluability and its implications for joint-separate evaluation reversals and beyond. In Kahneman, D. and Tversky, A., editors, *Choices, Values and Frames*, pages 543–565. Cambridge University Press.

Huttenlocher, J., Hedges, L., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3):352–376.

Insko, C., Schopler, J., Graetz, K., Drigotas, S., Currey, D., Smith, S., Brazil, D., and Bornstein, G. (1994). Interindividual-intergroup discontinuity in the prisoner's dilemma game. *Journal of Conflict Resolution*, 38(1):87–116.

Insko, C., Schopler, J., Hoyle, R., Dardis, G., and Graetz, K. (1990). Individual-group discontinuity as a function of fear and greed. *Journal of Personality and Social Psychology*, 58(1):68–79.

Iyengar, S., Wells, R., and Schwartz, B. (2006). Doing better but feeling worse: Looking for the best job undermines satisfaction. *Psychological Science*, 17(2):143–150.

Juslin, P., Winman, A., and Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, 114(3):678–703.

Kahn, B. (1995). Consumer variety-seeking among goods and services:: An integrative review. *Journal of Retailing and Consumer Services*, 2(3):139–148.

Kahneman, D., Fredrickson, B., Schreiber, C., and Redelmeier, D. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405.

Kahneman, D. and Thaler, R. (2006). Anomalies: Utility maximization and experienced utility. *Journal of Economic Perspectives*, 20(1):221–234.

Kahneman, D., Wakker, P., and Sarin, R. (1997). Back to bentham? explorations of experienced utility. *The Quarterly Journal of Economics*, 112(2):375–405.

Kameda, T., Tindale, R., and Davis, J. (2003). Cognitions, preferences, and social sharedness: Past, present, and future directions in group decision making. In Schneider, S. and Shanteau, J., editors, *Emerging perspectives on judgment and decision research*, pages 458–485. Cambridge University Press.

Kocher, M. and Sutter, M. (2005). The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *Economic Journal*, 115(500):200–223.

Kocher, M. and Sutter, M. (2007). Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments. *Empirica*, 34(1):63–88.

Kugler, T., Bornstein, G., Kocher, M., and Sutter, M. (2007). Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic Psychology*, 28(6):646–657.

Larrick, R. and Soll, J. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52(1):111–127.

Laughlin, P. and Ellis, A. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, 22(3):177–189.

Lee, P. and Brown, N. (2004). The role of guessing and boundaries on date estimation biases. *Psychonomic Bulletin & Review*, 11(4):748–754.

Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics*, 118(4):1209–1248.

Loewenstein, G., Thompson, L., and Bazerman, M. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social psychology*, 57(3):426–441.

Loomes, G. (2005). Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. *Experimental Economics*, 8(4):301–323.

Loomes, G., Orr, S., and Sugden, R. (2009). Taste uncertainty and status quo effects in consumer choice. *Journal of Risk and Uncertainty*, 39(2):113–135.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368):805–824.

Loomes, G. and Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, 53(2):271–282.

Luhan, W., Kocher, M., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12(1):26–41.

Manski, C. (1977). The structure of random utility models. *Theory and Decision*, 8(3):229–254.

March, J. (1996). Learning to be risk averse. *Psychological Review*, 103(2):309–319.

McCallum, D., Harring, K., Gilmore, R., Drenan, S., Chase, J., Insko, C., and Thibaut, J. (1985). Competition and cooperation between groups and between individuals. *Journal of Experimental Social Psychology*, 21(4):301–320.

Mellers, B. (2000). Choice and the relative pleasure of consequences. *Psychological Bulletin*, 126(6):910–924.

Mellers, B. and McGraw, A. (2001). Anticipated emotions as guides to choice. *Current Directions in Psychological Science*, 10(6):210–214.

Mellers, B., Schwartz, A., Ho, K., and Ritov, I. (1997). Decision affect theory. *Psychological Science*, 8(6):423–429.

Mellers, B., Schwartz, A., and Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, 128(3):332–345.

Morris, P. (1974). Decision analysis expert use. *Management Science*, 20(9):1233–1241.

Morris, P. (1977). Combining expert judgments: A bayesian approach. *Management Science*, 23(9):679–693.

Rauhut, H. and Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, 55(2):191–197.

Read, D., Antonides, G., Van den Ouden, L., and Trienekens, H. (2001). Which is better: simultaneous or sequential choice? *Organizational Behavior and Human Decision Processes*, 84(1):54–70.

Read, D. and Loewenstein, G. (1995). Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices. *Journal of Experimental Psychology: Applied*, 1(1):34–49.

Read, D. and Van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, 76(2):189–205.

Riis, J., Loewenstein, G., Baron, J., Jepson, C., Fagerlin, A., and Ubel, P. (2005). Ignorance of hedonic adaptation to hemodialysis: A study using ecological momentary assessment. *Journal of Experimental Psychology: General*, 134(1):3–9.

Schelling, T. (1980). *The strategy of conflict*. Harvard university press.

Schkade, D. and Kahneman, D. (1998). Does living in california make people happy? a focusing illusion in judgments of life satisfaction. *Psychological Science*, 9(5):340–346.

Schopler, J., Insko, C., Drigotas, S., Wieselquist, J., Pemberton, M., and Cox, C. (1995). The role of identifiability in the reduction of interindividual-intergroup discontinuity. *Journal of Experimental Social Psychology*, 31(6):553–574.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and Lehman, D. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178–1197.

Sen, A. (1977). Social choice theory: A re-examination. *Econometrica*, 45(1):53–89.

Sherman, D. and Cohen, G. (2002). Accepting threatening information: Self-affirmation and the reduction of defensive biases. *Current Directions in Psychological Science*, 11(4):119–123.

Simonson, I. (1990). The effect of purchase quantity and timing on variety-seeking behavior. *Journal of Marketing Research*, 27(2):150–162.

Smith, J. and Winkler, R. (2006). The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322.

Soll, J. and Larrick, R. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):780–805.

Steiner, I. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, 11(4):273–283.

Steiner, I. (1972). *Group process and productivity*. Academic Press New York.

Stewart, T. (2001). Improving reliability of judgmental forecasts. In Armstrong, J., editor, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pages 81–105. Kluwer Academic Publishers.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Random House of Canada.

Tajfel, H. and Turner, J. (1979). An integrative theory of intergroup conflict. In Austin, W. and Worchel, S., editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks Cole.

Thurstone, L. (1945). The prediction of choice. *Psychometrika*, 10(4):237–253.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Ubel, P., Loewenstein, G., and Jepson, C. (2005). Disability and sunshine: Can hedonic predictions be improved by drawing attention to focusing illusions or emotional adaptation?. *Journal of Experimental Psychology: Applied*, 11(2):111–123.

Van den Steen, E. (2004). Rational overoptimism (and other biases). *American Economic Review*, 94(4):1141–1151.

Vul, E. and Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

Wildschut, T. and Insko, C. (2007). Explanations of interindividual-intergroup discontinuity: A review of the evidence. *European Review of Social Psychology*, 18(1):175–211.

Wildschut, T., Pinter, B., Vevea, J., Insko, C., and Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, 129(5):698–722.

Wilson, T. and Gilbert, D. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, 14(3):131–134.

Wilson, T., Wheatley, T., Meyers, J., Gilbert, D., and Axsom, D. (2000). A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 78(5):821–836.

Winkler, R. (1981). Combining probability distributions from dependent information sources. *Management Science*, 27(4):479–488.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Review of Psychology*, 51(1):539–570.

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, 13(2):75–78.

Yaniv, I., Choshen-Hillel, S., and Milyavsky, M. (2011). Receiving advice on matters of taste: Similarity, majority influence, and taste discrimination. *Organizational Behavior and Human Decision Processes*, 115(1):111–120.

Yaniv, I. and Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281.