ON ESTIMATE AGGREGATION

Studies of how decision makers aggregate quantitative estimates in three different cases

Piya Sereevinyayut

TESI DOCTORAL UPF / 2013

DIRECTOR DE LA TESI

Dr. Robin M. Hogarth

DEPARTAMENT D'ECONOMIA I EMPRESA

Universitat Pompeu Fabra Barcelona

To my family and friends

Acknowledgements

This dissertation is long in the making, filled with every emotion from paralyzing dreads of failure to heart-pounding joys of discovery. It has been a journey full of self-doubts and self-discoveries that I am now glad to embark on.

I would like to extend my utmost gratitude towards my advisor, Prof. Robin Hogarth. Without his great patience I would have given up long before. His advices and supports, of which I made use regretfully not to the fullest, have been infinitely crucial in each step in helping me reach this finish line.

I am grateful for all comments from the UPF community, especially an assistance from Prof. Albert Satorra for his advices in the issues of statistical methods. I would like also also mention a few friends who brightened the path along this journey for me. I am thankful for decades of friendship with Yuwapat Vasunirachorn who was always there when I needed someone to talk to, for Pataporn Sukontamarn who never gets tired from giving me encouragements, for Kanda Naknoi whom I can always look up too, and for friendly supports from Elena Reutskaja and Javier Palacio.

Abstract

This dissertation examines how people aggregate quantitative advices to reach their own estimates. Each chapter explores a different situation that could affect how advices are evaluated, and consequentially how advices will be combined. The first chapter demonstrates that people measure advices' extremity degrees by anhoring upon the advice set's median. It also shows that, unlike multiplicative scaling, additive scaling of advices affects how outliers are perceived. The second chapter deals with advices that are obtained serially. The results reveals that whether people execute the aggregation sequentially or only once at the end of the series affects how an outlier in the series is detected and combined. The third chapter studies how people revise their own estimates with advices of others, and finds that people revise more if they appear a dissensus. Consequentially having multiple advices can attenuate of the effect of egocentricity and improve accuracy of revisions compared to having only a single advice.

\

Resum

Aquesta tesi estudia com les persones agreguen consells quantitatius per arribar a les seves pròpies estimacions. Cada capítol explora una situació diferent que podria afectar com s'avaluen els consells, i en conseqüència com es combinen aquests consells. El primer capítol demostra que les persones mesuren els graus extrems dels consells per ancoratge a la mediana del conjunt de consells. També es mostra que, en comptes d'una escala multiplicadora,l' escala additiva dels consells afecta a com es perceben els valors atípics. El segon capítol tracta de consells que s'obtenen en sèrie. Els resultats revelen que si les persones executen l'agregació seqüencialment o només una vegada al final de la sèrie, afecta a com es detecten i es combinen els valors atípics en la sèrie. El tercer capítol estudia com les persones revisen les seves estimacions a partir consells dels altres, i es troba que les persones revisen més si es troben en un dissens. Conseqüentment, tenir consells múltiples pot atenuar l'efecte d'egocentrisme i millorar la precisió de les revisions si es compara en tenir només un únic consell.

Foreword

We all solicit advices and opinions to aid our decision making, but can those advices help us improve our decisions? Do different environments and contexts that advices are presented result in different decisions? Or do we even use those advices at all? This dissertation consists of three self-contained chapters, each examining the same issue of how decision makers aggregate quantitative advices, such as sales estimates or market forecasts, but each under different decisional contexts that could affect how advices are evaluated and combined. After receiving an estimate, the first question one might ask is "Is it believable?" Obviously, if the answer is no then one would discard it or at least reduce its weight in the aggregate. However this significantly depends on one's ability to identify the believability degree of a received estimate. Two chapters in this dissertation deal with the topic of extreme estimates, or outliers, specifically whether people could detect it; the first chapter examines the differential effect of data scaling in how outliers are perceived, and the second chapters examines how task processes could intervene in outlier detection when estimates are presented sequentially. But it is not just extreme estimates that one would be likely to ignore. We are often so egocentric that we use advices just to confirm, rather than correct or revise our initial ideas. The last chapter in this dissertation examines whether with multiple advices, one's egocentricity could be reduced, so one would take advices and revise one's own initial opinion accordingly.

The first chapter titled "The Outlier Identification, Scaling Effects, and Forecast Aggregation" presents the results from two laboratory experiments that examine how people define and detect outliers, which demonstrate that experimental participants based their judgments of outlier extremity on comparative distance from the median of the forecasts to be aggregated. Perceptions of experimental participants with regard to outliers were also affected by additive scaling (where a constant is added to all forecasts) but not by multiplicative scaling (where the data are multiplied by a constant). At the end, participants' aggregation strategies were generally heavily anchored on the median of data observed, which readily reduced the impact of outliers in aggregations.

The second chapter titled "Effects of outlier appearance order in aggregation of short forecast sequences" shows that, from the results of two laboratory experiments, aggregation processing modes also plays a role in whether people could detect outliers. Specifically when participants produced and updated their estimates with each new

forecast along a sequence, they missed an outlier if it appeared early but their final estimates showed no influence of that outlier, but an outlier that appeared last got detected but still ended up with a substantial weight in the final aggregation; and when participants were made to recall earlier forecasts in a sequence and to produce only one single estimate at the end of each sequence, they duly realized forecast extremity, but outliers still exerted a weighty influence over the final aggregation regardless of order of appearance.

The third and final chapter titled "Estimate revision with multiple advices" explores egocentricity in opinion revision. Results from the laboratory experiment demonstrate that people will choose to revise more when they find their opinions to be outside a consensus. In fact, further analyses show that this consensus-dissensus category is a valid cue for an accuracy judgment. The second laboratory experiment in this chapter looks at whether concerns for rankings would make people make even more use of advices as previous research argues, an the results did not support this hypothesis. A simulation study using the data collected from two experiments suggests that having multiple advices, and its consequential revise-if-dissensus heuristic, can improve accuracy of revisions that decision makers may choose compared to having only a single advice, especially that a single advice is often egocentrically ignored.

Contents

List of figures

List of tables

1. Outlier Identification, Scaling Effects, and Forecast Aggregation

Abstract.

Faced with several forecasts, how do people define and detect outliers? Moreover, how do they take account of identified outliers when aggregating such forecasts? Two experiments examine these questions and suggest that experimental participants based their judgments of outlier extremity on comparative distance from the median of the forecasts to be aggregated. Participants' perceptions of outliers were also affected by additive scaling (where a constant is added to all forecasts) but not by multiplicative scaling (where the data are multiplied by a constant). In general, participants' aggregation strategies were heavily anchored on the median of data observed.

1.1 Introduction

Rarely is an important decision taken without asking for a second opinion (Harvey & Fischer, 1997). However, any expert can be biased, misinformed or under-informed (Hogarth & Makridakis, 1981; Goodwin & Wright, 1994; Webby & O'Connor, 1996); in these situations additional opinions can reduce or offset such biases. Research has shown that combining forecasts from multiple sources results in improved accuracy (Einhorn, 1972; Dawes & Corrigan, 1974; Doyle & Fenwick, 1976; Libby & Blashfield, 1978; Makridakis et al., 1982; Yaniv & Hogarth, 1993; Armstrong, 2001, Winkler & Clemen, 2004). In particular, unit weighting is an efficient aggregation scheme (Winkler, 1971; Newbold & Granger, 1974; Einhorn & Hogarth, 1975; Hogarth, 1978; Libby & Blashfield, 1978; Clemen & Winkler, 1986; Lawrence et al., 1986; de Menezes et al., 2000), significantly improving forecast accuracy without requiring additional information.

But the use of a simple average is often underestimated (Sniezek & Henry, 1989; Larrick & Soll, 2006). Instead, with multiple forecasts at hand decision makers (DMs) often attempt to evaluate advisors based on certain characteristics and to aggregate their forecasts accordingly. For example, Maines (1996) showed that experimental participants put more weight on forecasts by advisors whose past forecasts had been more accurate. They also adjusted their estimates according to their views of advisors'

biases, that is, whether they were optimistic, neutral or pessimistic. Experiments by Budescu et al. (2003) demonstrated that participants weighted advisors' opinions according to the accuracies of their previous forecasts as well as the amount of information they might have used. Yaniv and Foster (1995) allowed advisors to be imprecise in their opinions and found that experimental participants gave greater weight to more precise opinions. In fact, when the information about advisors is accurately perceived, subjective weighting can improve aggregation beyond simple averaging (Ashton & Ashton, 1985; Flores & White, 1989; Fischer & Harvey, 1999). However there is also a possibility that DMs might misuse such information, and as a consequence produce biased aggregations (Soll, 1999; Larrick & Soll, 2006).

In many situations, however, and particularly in the one-shot case, the only information available is the forecasts or opinions themselves. Here, a simple average seems to be a sensible choice as forecast evaluation is not possible. Even so, given the human penchant for pattern and order, a DM could try to construct a pattern just from the forecasts, such as consensus and outlier, and evaluate each accordingly. When confronted with outlying opinions, some research has shown that DMs tend to take the median as the aggregate (Yaniv, 1997; Harries et al., 2004). However, a "knee-jerk" dismissal of an extreme forecast could be an overreaction. A forecast could seem like an outlier because its provider does not participate in so-called "herding". In this case, an outlier can provide additional information and be a good hedge in aggregation. It has been shown, for example, that bold forecasts often come from experienced analysts who incorporate more information in their forecasts, whilst young and inexperienced analysts try to avoid the negative consequences of forecast errors and revise their forecasts to conform to a consensus (Chevalier & Ellison, 1999; Hong et al, 2000; Clement & Tse, 2005). On the other hand, a bold forecast could also result from manipulation that is not information-based (Lamont, 2002).

In this chapter, I first asked how a DM confronted with several forecasts judges one to be an outlier. Results from two laboratory experiments indicated that participants were sensitive to the comparative distance between an extreme forecast and the other data points. At the same time, participants' perceptions were differentially affected by additive and multiplicative transformations (i.e., rescaling) of the data. Specifically, they were seen to be sensitive to additive (adding a constant to all opinions, e.g. a company's revenue increase versus a company's total revenue) but not multiplicative changes (multiplying everything by a constant, e.g. a company's revenue in different

currencies). This means DMs' views on the likelihood distribution of a true value that was being forecast could be altered by using a different scaling in a presentation of data. Second, I investigated how outliers were treated in aggregating forecasts and found that the extent to which an outlier was perceived as deviant generally had no significant impact on how it was weighted. Instead, participants seemed to anchor their estimates around the sample median. The distribution of the estimates also suggested that the level of the data (additive scaling) played a role in attenuating subjective forecast variability.

1.2 Outlier perception

1.2.1 Measuring outlying degrees

Essentially identifying an outlier is to find a data point that is either "too large" or "too small" relative to those with which it is grouped, a data point that is "distant" from the other data points in the set. This implies that the degree of outlier extremity, i.e. the likelihood that a data point is a true outlier, can be evaluated by the distance between that data point and an anchor that is representative of the rest of the data in the set. Denote this distance as Absolute Distance (AD) which is expressed as following: in the forecast set $j, X_j = \{x_j^1, ..., x_j^n, ...x_j^N\}$, AD of x_j^n is

$$AD(x_j^n) = | x_j^n - \alpha(X_j) | \tag{1.1}$$

where $\alpha(\cdot)$ is the anchor against which a data point is measured. But the definition also implies that a "true" outlier will lie at a greater distance away in comparison with other data points. That is, an outlier expectedly has an AD larger than the ADs of all the other data points in the set. Hence, the degree of outlier extremity depends on Comparative Distance (CD), where CD of x_j^n is defined as

$$CD(x_j^n) = \frac{AD(x_j^n)}{\frac{1}{N}\sum_k^N AD(x_j^n)} \tag{1.2}$$

As the distance AD measures how far a data point is from the rest, the anchor should represent the other data in the set. I consider three candidates for anchor that a DM might use: (1) the median, as discussed above, DMs often use the median as a choice for

aggregation when facing an outlier and thus may also use the median as an anchor in deciding whether there is an outlier; (2) the trimmed mean, i.e. the mean of the other forecasts; and (3) the nearest neighbor (NN), i.e. the distance to another datum whose value is the closest.

Among the three candidates for anchor, the trimmed mean requires the most computation and so is not readily accessible by a DM. On the other hand, both the median and NN, requiring only data sorting, are more likely candidates, especially if the sample size is large. In fact, using a sample size of 10, Collett and Lewis (1976) examined if either the trimmed mean or NN was used as an anchor, and found the impact of the latter to be significant but not that of the former.[1]

Still, the extent to which a DM perceives an outlier as extreme can depend on perspective. The same distance might seem normal to a DM under one set of circumstances but look inordinate otherwise. Essentially, outlier identification relies on a DM's presumption of the variability of the forecasts in consideration. But people's intuitive inferences about variance are not infallible (Peterson & Beach, 1967). Specifically, people generally presume that larger numbers imply larger variance. As a consequence, increasing the level (L) of the data, i.e. the overall magnitude of the data set such as sample mean or median, can attenuate a DM's perception about the size of the variance of the forecasts, despite the fact that the statistical variance does not change at all. For example, in experiments by Lathrop (1967), participants ranked cards with separated lines from high to low in variability. He found that the ranking was inversely proportional to the square root of the mean length of lines in the cards. Beach and Solak (1969) had experimental participants estimate a percentage of a certain number, and state the interval of acceptable estimated errors. They found that the widths of the intervals were proportional to the magnitude of the correct answers. In Lawrence and Makridakis's (1989) experiment, participants were asked to provide estimates and confidence intervals for time series data. It turned out that participants widened their confidence intervals when the data series demonstrated a trend, i.e. a change in the data level, both upward and downward. So the data level, which affects the perception of variability, might also be expected to affect the perception of outliers.

1.2.2 Differential effects of scaling

[1] The regression models compared in Collett and Lewis (1976) included NN divided by the data range, and trimmed mean divided by the standard deviation of non-outlier predictions.

From the discussion above, it is expected that adding a large number (constant) to the forecast set could affect outlier identification and consequently a DM's aggregate. I call this additive- or A-scaling. This type of scaling is not uncommon, for example, yearly sales forecasts could be communicated as the total sales of the coming year or as the year-to-year sales increase. Thus, two DMs receiving the same sales information but with different scaling could reach different aggregates. One might discard one forecast for being extreme and therefore take the median, while the other might consider all forecasts as equally valid and use the mean.

A forecast can also be defined using different units, for example yearly sales versus quarterly sales, sales per store versus sales per square foot. Neither is such multiplicative- or M-scaling rare. As an example, contrast the case of forecasts being expressed in terms of, say, Japanese yen as opposed to US dollars.[2] One might expect that since smaller units result in larger numbers (i.e., accounting in yen as opposed to dollars), larger absolute distances regardless of anchors could amplify the extent of perceived outlier extremity. And because of this, the final aggregate of the DM might de-emphasize the role of that data point. Conversely, the smaller unit implies a higher data "level" (i.e., larger numbers) and, as a consequence, a DM could also include a "true" outlier in the aggregate. Thus, the outcome of M-scaling depends on which of these two opposing effects is stronger.

To study the impact of scaling, I will compare DMs' perceptions of a data point regarding its likelihood to be an outlier before and after that data point is transformed by each of the two scaling types. Furthermore I will examine how the above hypothesized factors – AD, CD and L – affect outlier perception, in similar fashion to Collett and Lewis (1976), using the following two logistic models,

Model 1: $$\lambda(x_j^n) = \alpha + \beta \cdot CD(x_j^n) + \gamma \cdot AD(x_j^n) + \delta \cdot L \qquad (1.3.1)$$

Model 2: $$\lambda(x_j^n) = \alpha + \beta \cdot CD(x_j^n) + \gamma \cdot \frac{AD(x_j^n)}{L} \qquad (1.3.2)$$

where $\lambda(x_j^n)$ are the log-odds that x_j^n is seen as an outlier. One important difference between the two models is that Model 2 does not assume an independent effect of data

[2] The yen-dollar exchange rate is approximately 77 to 1 (Jan, 3, 2012).

level (L); instead it enters the equation as a ratio involving the absolute distance, i.e. AD/L. Thus, if Model 2 is valid, there should be no M-scale effect.

1.3 A-scale experiment

1.3.1 Methods

Procedures. The experiment was conducted in a laboratory in Barcelona in Spanish on personal computers. Participants were given the instruction sheets which were also read out loud to them. The instructions stated that the experiment was about sales forecasting of one supermarket chain which operated stores of various sizes across the southeastern United States. This supermarket held annual meetings where senior managers and executives from different departments gathered. At these meetings, the executives gave the forecasts for the sales of the following year of the stores they supervised. The instructions further specified that the data came from the 2002 annual meeting hence the real sales were known and exact evaluations of the forecast accuracies were possible. Participants were informed that the experiment contained 18 rounds concerning sales of 18 different stores, 1 store per round. In each round they would be given a set of four forecasts about monthly sales in units of thousand U.S. dollars of one particular store randomly selected from a pool of forecasts given at the company's annual meeting. And there would be two tasks to perform per round.

In the first task, each participant was asked to bet on the accuracies of the four forecasts presented in that round, by distributing 100 tokens among forecasts. The payoffs followed a proper scoring rule: a bet of T tokens on a forecast incurred a cost of $T^2/100$ points; if that forecast turned out to be the most accurate of all four, the participant would win $2T$ points. The proper scoring rule implies that the optimal distribution of bets follows the subjective comparative probabilities that each forecast will be the most accurate. Hence the amount bet on an outlier was assumed to reflect its perceived outlier extremity (inversely), i.e., the lower the amount bet on a forecast, the more likely it was assumed to be an outlier. Participants were told explicitly that they should distribute tokens on forecasts according to the likelihood to be the most accurate, and that a forecast they deemed more likely to be the most accurate should receive more tokens. The maximum net gain for this task was 100 points.

In the second task, participants estimated the sales themselves where they could base it on the four forecasts in any way they liked. The payoff depended on the accuracy of their estimates. Specifically, if their answer deviated below 1 unit from the realized (i.e. true) value they would win 100 points, if the deviation was at least 1 but below 3 units they would receive 60 points, and if the deviation was at least 3 but below 10 units the winning was 10 points. Deviations of 10 or more units earned them no points. With this payoff structure, the best strategy was to give the most accurate estimate possible.

The two tasks and the payoff schemes described above were explained to participants. The instructions contained one example of a hypothetical round with four forecasts, hypothetical bets and a hypothetical estimate, and where the calculations of the payoffs were shown in detail. Participants were also told explicitly that the amount bet on each forecast should correspond to the likelihood that it would be the most accurate of all four, to match the proper scoring rule. The experiment started with two practice rounds to familiarize participants with the interface of the experiment. For these two practice rounds, they received feedback that included realized sales and detailed calculations of the payoffs based on their bets and estimates. The exchange rate for payoffs was 200 points for 1 Euro.

Stimuli. First the non-outlier forecasts were randomly generated following a normal distribution with a mean of 85 and a standard deviation of 17. This would be the base, i.e. pre-scaled, forecasts. They were limited to lie within two standard deviations from the mean to assure that later-planted outliers would be the most extreme forecasts. These forecasts were divided into 20 sets of three forecasts.

Table 1.1. *Matching matrix of forecast sets and participants in each round*

	Participant 1	**Participant 2**	**...**	**Participant 19**	**Participant 20**
Round 1	set 1	set 2	...	set 19	set 20
Round 2	set 2	set 3	...	set 20	set 1
:	:	:	:	:	:
Round 18	set 18	set 19	...	set 16	set 17

The number of sets matched the number of participants in the experiment. The matching between participants and forecast sets was as shown in Table 1.1 so that (1) for any

single participant no two rounds used the same set, (2) for any single round, no two participants used the same set, and (3) the same 20 sets were used in all rounds. There were three levels of outlier, i.e. 2, 3 or 4 standard deviations (s.d.'s) and outliers were either left (-) or right (+) of the population mean. One outlier was included in the forecast set of each round, as shown in Table 1.2. The order in which the outliers appeared on screen (in a vertical row of four forecasts) was random.

Table 1.2. *Types of forecast outlier in each round*

Rounds	Outlier type	Rounds	Outlier type
1, 7, and 13	- 2 s.d.	**4, 10, and 16**	+ 2 s.d.
2, 8, and 14	- 3 s.d.	**5, 11 and 17**	+ 3 s.d.
3, 9, and 15	- 4 s.d.	**6, 12, and 18**	+ 4 s.d.

I considered three levels of scaling manipulation, as described in Table 1.3. For the Base level the pre-scaled forecasts were used; for the A1 level I increased the data level 3 times that of the Base while maintaining the same spread and variation by adding 170 to all stimuli (both outliers and non-outliers); for the A2 level I raised the data level to 7 times that of the Base but kept the same spread and variation by adding 510 to all forecasts. The order of rounds was randomized for each individual participant.

Table 1.3. *Forecast manipulation in each round and its key statistical properties*

	Scaling level	Manipulation	Statistical properties	
			Mean	Std. var.
Rounds 1 - 6	Base	n/a	85	17
Rounds 7 - 12	A1	adding 170	255	17
Rounds 13 - 18	A2	adding 510	595	17

Participants. Twenty participants aged between 18 and 22 years took part in this experiment. Twelve were female, 8 were male. They were recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the

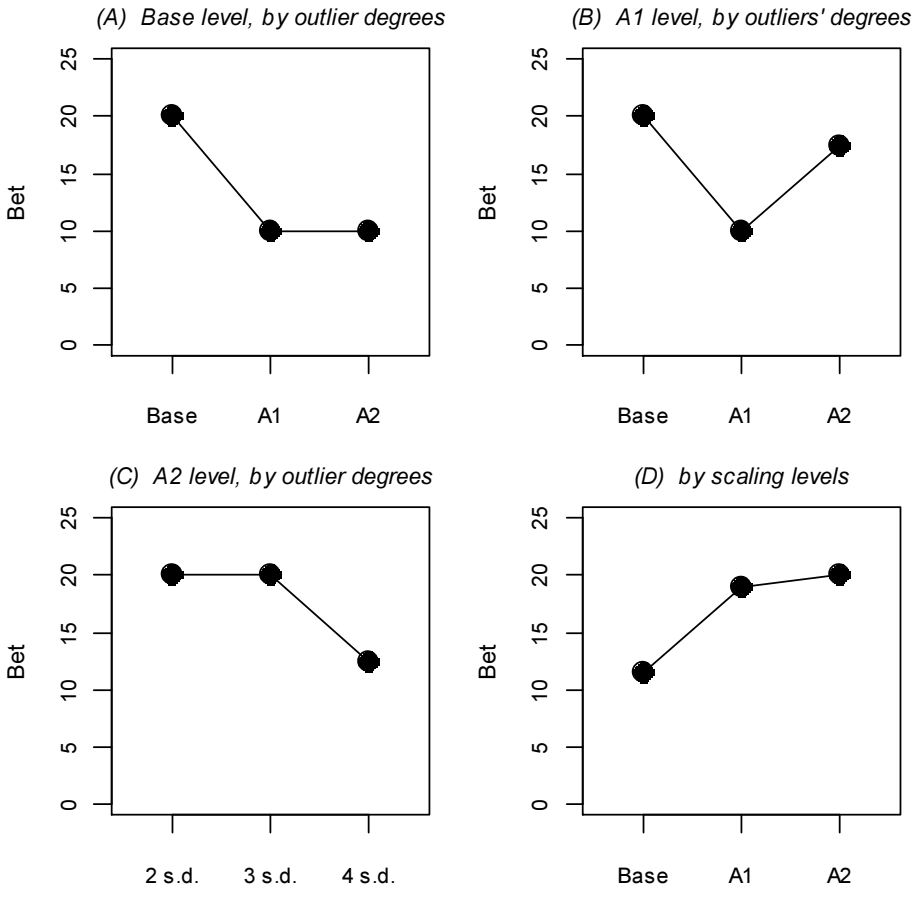
experimental laboratory. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 10.50 Euros.

1.3.2 Results

1.3.2.1 Outlier perception

Scaling and outlier perception. The median bets for different levels of scaling are depicted in Figure 1.1. In the Base level (see Panel A), as expected, more extreme outliers received higher bets (Medians = 20, 10, and 10 for outliers of 2, 3, and 4 s.d. respectively), this effect of outlier degrees was significant ($p<.05$, Kruskal-Wallis test), especially that outliers at 2 s.d. received significantly lower bets than outliers of 3 and 4 s.d. ($p<.01$, Wilcoxon test).

Figure 1.1. *Median bets placed on outliers (A-scale)*



While in the A1 level (see Panel B), the similar trend did not prevail (Medians=20, 10, and 17.5 for outliers of 2, 3, and 4 s.d. respectively), and the difference of bets the three

outlier degrees received was not significant (p≈.11, Kruskal-Wallis test). Although in the A2 level (see Panel C) we can see that participants placed higher bets on less extreme outliers (Medians=20, 20, and 12.5 for outliers of 2, 3, and 4 s.d. respectively), the effect of outlier degrees was not significant (p≈.22, Kruskal-Wallis test). Even the difference between bets for outliers at 4 s.d. versus bets for outliers at 2 and 3 s.d. only approached significance (p<.10, Wilcoxon test).

Overall (see Panel D) higher scaling levels resulted in higher bets placed on outliers (Medians=11, 19, and 20 for the Base, A1, and A2 levels respectively). While the effect of outlier levels only approached significance (p<.10, Kruskal-Wallis test), outliers at the Base level received significantly lower bets than those at the A1 and A2 levels (p<.05, Wilcoxon test).

Factors affecting outlier perception. The analysis followed the two models described in Equations 1.3.1 and 1.3.2 and was only applied to the planted outliers. Since the quantity of tokens placed as a bet was used as a proxy for outlier perception, I considered that a participant saw an outlier as an outlier if the bet placed on it was below a certain threshold. The median bet (on planted outliers) of 15 was chosen as this threshold. Sample medians were used as the data levels because they represented the central values of the forecast sets without including the effect from the size of planted outliers, and because, as we will see later, participants tended to anchor their judgments in the experiment on sample medians. The design of the experiment was that each participant gave responses to multiple rounds of tasks. Due to these correlated responses, the method of Generalized Estimating Equations (GEE) was used. The results of the regressions are shown in Table 1.4.

Table 1.4. *Logistic regressions of outlier identification likelihood (A-scale)*

	Trimmed-mean anchor		**Median anchor**		**NN anchor**	
	Model 1	**Model 2**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
Intercept	- 3.114*	- 3.687**	- 2.655***	- 2.806***	- 2.456***	- 2.633***
CD	0.465	1.030*	0.258	0.445**	0.336	0.427**
AD	0.024'	n/a	0.093'	n/a	0.011	n/a
L	- 0.000	n/a	- 0.000	n/a	- 0.000	n/a
AD/L ratio	n/a	1.037**	n/a	1.177*	n/a	1.181*

*Significant codes: ' p-value<.1, * p-value <.05, ** p-value<.01, *** p-value<.001*

The coefficients show the expected signs, that is, the larger distances and the lower data levels increased the likelihood of the extreme value being seen as an outlier. With Model 1, only the coefficient of the absolute distances approached significance, while the data levels had no effect, contrary to what the findings in previous research would have suggested (Lathrop, 1967; Beach & Solak, 1969; Lawrence & Makridakis, 1989).

To compare overall fits of these two models, because of their being non-nested and because of within-participant dependence of responses, I used the nonparametric test suggested by Clarke (2003, 2007). The tests confirmed that Model 2 fitted significantly better across all anchors ($p<.01$ with trimmed mean, $p<.001$ with medians, and $p<.001$ with NN).[3] These suggest that the data level entered participants' consideration together with the absolute distance as a single ratio. This outcome anticipated the absence of M-scale bias, which I will examine in the next experiment. Clarke's test was also used to compare the fits under different anchors. I found that the trimmed-mean anchor performed the worst ($p<.01$ when compared to either median or NN). NN fitted slightly better than median but not significantly ($p\approx.40$).

Figure 1.2. *Median predicted probabilities of outlier identification (A-scale)*

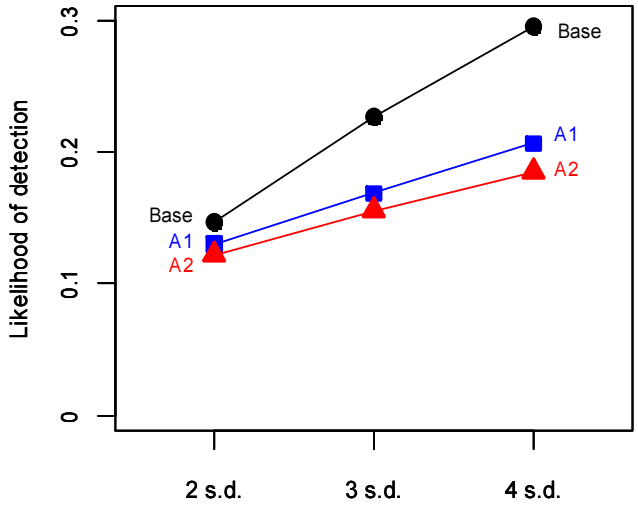


Figure 1.2 plots the median predicted likelihoods of outlier identification by Model 2 with NN as an anchor under different scaling and outlier levels. We can clearly see an

[3] I also performed regressions using a modified Model 1 with no absolute distance as a variable. The results were again that the data level was not significant in any of the anchor choices. In addition, from Clarke's tests, Model 2 fitted significantly better ($p<.001$ in all anchor choices).

A-scale effect from the apparent separation between Base and the other two scaling levels, where the gap widens as the outlier becomes more extreme.

1.3.2.2 Outlier perception and forecast aggregation

Aggregate anchor. Prior research has argued that "take-the-median" is a likely heuristic when a DM faces an outlier. But we have shown that A-scaling can affect what people perceive as being an outlier. The principal interest in this section was to examine whether the effect of A-scaling extended to how a DM chose an aggregate. As the main concern was in the judgmental weight accorded to outliers in aggregation, participants' estimates were framed as departures from the median *towards* an outlier.

Since the forecast sets in the experiment were all different, some standardization was required. I chose to measure deviations in units of the (outlier-excluded) mean deviation from the (outlier-included) median.[4] If an estimate had incorporated an outlier, this measurement of distance would not dilute the footprint of that outlier. The figures of the standardized departures from the median (SDM) in Table 1.5 show that participants' estimates were anchored around the median. The location tests (Wilcoxon tests) could not reject that the median was the center of aggregation in any of scaling levels and outlier degrees.

Table 1.5. *Median standardized departures from the median (A-scale)*

	Base scale	**A1 scale**	**A2 scale**
2 s.d.	0.13	- 0.01	0.05
3 s.d.	- 0.05	- 0.30	0.25
4 s.d.	0.14	0.58	- 0.29
overall	0.13	0.00	0.04

Outlier's aggregation weight. To examine the relation between outlier perception and aggregation, I performed median regressions (quantile regressions with

[4] For example, if the forecast set is {1, 4, 6, 25}, the median of the whole set that includes the outlier {25} is 5, and the deviations from the medians for each forecast are 4, 1, 1, and 20 respectively. Hence, the outlier-excluded mean is (4+1+1)/3 = 2. If the estimate from a participant is 8, then SDM is calculated as (8-5)/2 = +1.5. If the estimate is 3, then SDM is (3-5)/2 = -1.

τ=0.5) of SDM on bets that outliers received. Regression coefficients were obtained assuming working independence among responses both between- and within-participant, while for statistical inference within-participant correlations were simulated via block bootstrapping. The results (Table 1.6) showed that the coefficients of outlier aggregation weights were greater in higher scaling levels, but only significant at A2. Thus we could conclude that when a DM sees an outlier as an outlier, the heuristic is to take the median. But when an outlier is not identified, the DM is more comfortable including all forecasts in the combination, weighting them according to their perceived chances of being accurate.

Table 1.6. *Median regressions of aggregation scheme on outlier perception (A-scale)*

	Base	**A1**	**A2**
Intercept	- 0.135	- 0.430	- 0.312*
Bet	0.018′	0.028′	0.027*

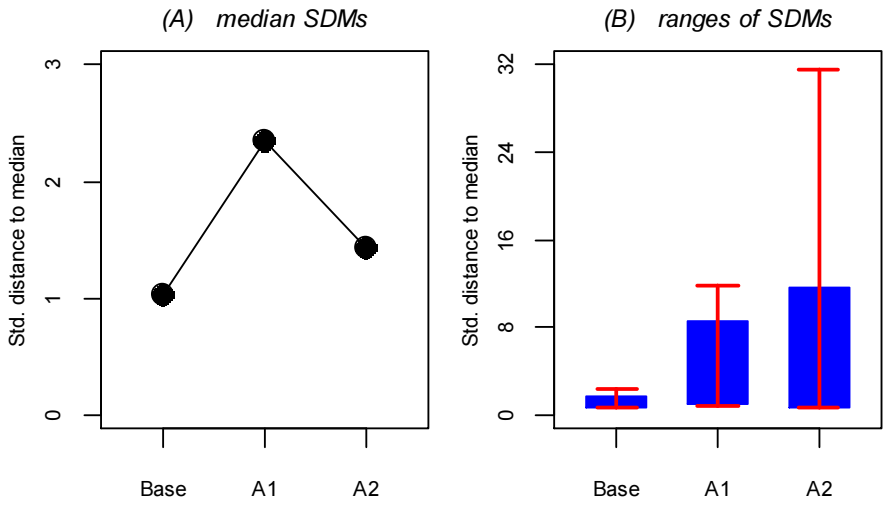
*Significant codes: ′ p-value<.10, * p-value <.05*

Subjective variability. Assume that, to produce an aggregate forecast a DM starts by considering the sample median. But then how much should the ensuing estimate deviate, either towards or away, from an extreme forecast? It is not unlikely that the DM's estimate will deviate far from the median if the forecasts are seen to involve high variability, and the opposite when low variability is perceived. Since high data level has been shown to attenuate the perception of variability (Lathrop, 1967; Beach & Solak, 1969; Lawrence & Makridakis, 1989), I expected that a higher A-scale level, seen earlier to obscure the extremeness of outliers, would result in a larger absolute SDM.

Figure 1.3 Panel A depicts median absolute SDMs which did not exhibit the expected trend (Median=1.03, 2.35, and 1.42 for the Base, A1, and A3 levels respectively). But SDMs from the Base level were significantly lower than both those from the A1 level ($p<.01$, Brunner-Munzel tests), and those from the A2 level ($p<.05$, Brunner-Munzel tests). While in Panel B, we can see that SDM ranges, which represented the span of participants' wandering away from the medians when computing the estimates, expanded upward with higher levels of A-scaling. After

adjusting for median locations, dispersion tests could not reject either that the A1 level had a larger span than the Base level (p<.001, Ansari-Bradley test), or that the span of A2 level exceeded that of the A1 level (p=.05, Ansari-Bradley test).

Figure 1.3. *Medians and ranges of standardized departures from the median (A-scale)*



Note: In Panel B, rectangular bars cover the middle-quintile, and lines cover the middle-quartile ranges.

1.3.2.3 Discussion

We have seen from this experiment that additive scaling led to attenuation in participants' perception of forecasts' extremeness, and consequently participants became more likely to deem an outlier as normal. The regression analyses revealed that when participants evaluated the extremity of a forecast by the distance between that forecast and others in the same set, it was done in a direct comparison to the overall level of forecasts. Since additive scaling raises the forecast level but keeps the distance between forecasts intact, a forecast in a set that is additively scaled up will look less distant, thus more likely. This also implies that multiplicative scaling will have no effect on outlier detection, which I will examine in the next experiment. In terms of aggregation, while prior research has argued that "take-the-median" is a likely heuristic when a DM faces an outlier, this experiment demonstrated that, regardless of whether outliers were detected or not, participants anchored their aggregation around medians. Furthermore, the examination of distributions of estimates from participants suggested that the additive-scaling might have raised subjective impressions of variability.

1.4 M-scale experiment

The previous section has demonstrated the effects in outlier perception caused by different levels of data. However the data levels were also used to benchmark the overall spread of forecasts, such that if we had scaled the data multiplicatively the same effects would not occur. To test this prediction, I conducted another experiment using the M-scaling manipulation.

1.4.1 Methods

Procedures. The same as in the A-scale experiment.

Stimuli. Stimuli in rounds 1 – 6 were maintained at the Base level. In rounds 7 – 12 all stimuli were multiplied by 3 (M1), and in rounds 13 – 18 by 7 (M2). So the data levels were comparable to those in the A-scale experiment (M1 with A1, and M2 with A2), and the only difference was the spread (Table 1.7).

Table 1.7. *Forecast manipulation in each round and its key statistical properties*

	Scaling	Manipulation	Statistical properties	
			Mean	Std. var.
Rounds 1 - 6	Base	n/a	85	17
Rounds 7 - 12	M1	Multiplying by 3	255	51
Rounds 13 - 18	M2	Multiplying by7	595	119

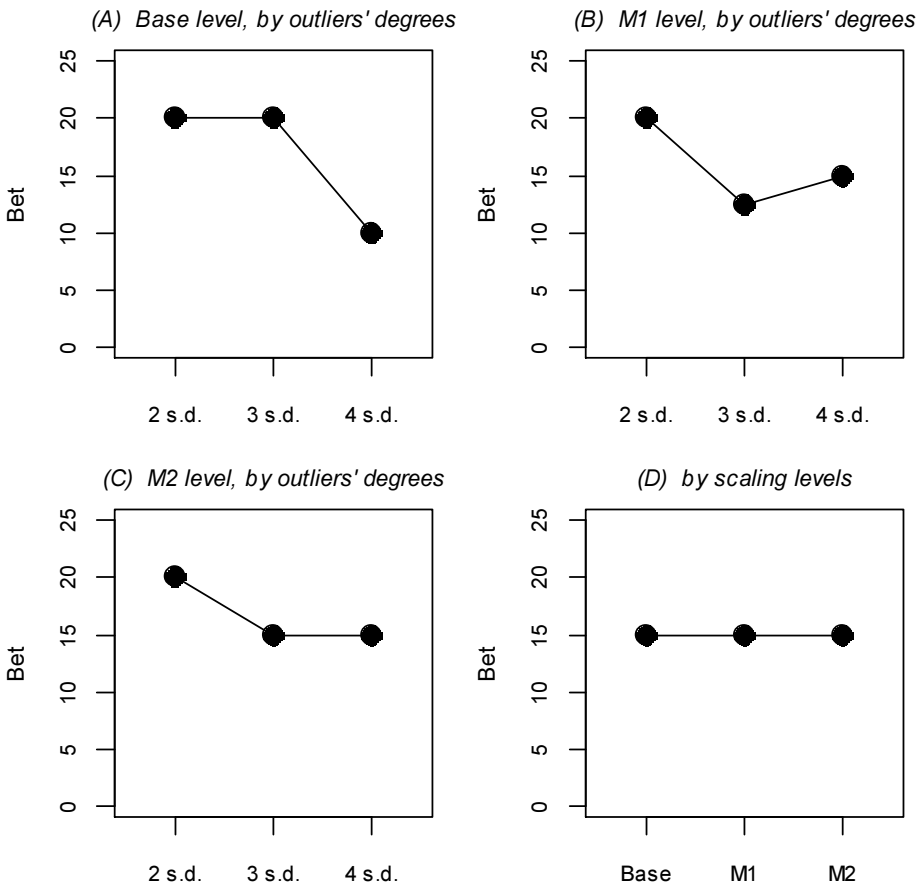
Participants. Twenty participants aged between 18 – 22 years took part in this experiment. Fourteen were female, 6 were male. They were recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the experimental laboratory. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 10.40 Euros.

1.4.2 Results

1.4.2.1 Outlier perception

Scaling and outlier perception. Figure 1.4 shows the results of bets placed on different outliers under different scaling levels. Overall, we can see the pattern that more extreme outliers received lower bets. However in the Base level (see Panel A), the effect of outlier degrees only approached significance (Medians=20, 20 and 10 for outliers at 2, 3 and 4 s.d. respectively; $p<.10$, Kruskal-Wallis test),even though bets that outliers at 4 s.d. received were significantly lower than bets that outliers at 2 and 3 s.d. received ($p<.05$, Wilcoxon test).

Figure 1.4. *Median bets placed on outliers (M-scale)*



In the M1 level (see Panel B), the effect of outlier degrees was not significant (Medians=20, 12.5 and 15 for outliers at 2, 3 and 4 s.d. respectively; $p\approx.40$, Kruskal-Wallis test), and bets that outliers at 2 s.d. received were not significantly lower than bets that outliers at 3 and 4 s.d. received ($p\approx.20$, Wilcoxon test). Similarly, in the M2 level (see Panel C) the outlier degree did not affect the levels of bets significantly

(Medians=20, 15 and 15 for outliers at 2, 3 and 4 s.d. respectively; p≈.18, Kruskal-Wallis test), but the difference between bets distributed to outliers at 2.s.d. versus to outliers at 3 and 4 s.d. approached significance (p<.10, Wilcoxon test). Overall (see Panel D) the effect of scaling level was not significant (p≈.18, Kruskal-Wallis test).In fact, on average all three scaling levels received the same bets (Median=15).

Factors affecting outlier perception. To examine the impacts of various factors on outlier perception involving the M-scale, I employed regression analysis based on Models 1 and 2 as above. The median bet used as a threshold for outlier perception was also 15. As shown in Table 1.8, the coefficients had the expected signs, that is, the likelihood of being seen as an outlier was higher with the larger distances and the lower data levels. The comparative distance was significant in both models and for all anchors; however the effect of the data levels and the absolute distances in both models were not significant.

Table 1.8. *Logistic regressions of outlier identification likelihood (M-scale)*

	Trimmed-mean anchor		**Median anchor**		**NN anchor**	
	Model 1	**Model 2**	**Model 1**	**Model 2**	**Model 1**	**Model 2**
Intercept	- 2.006*	- 2.259*	- 1.052*	- 1.120*	- 0.672	- 0.716'
CD	1.146*	1.343*	0.514*	0.577*	0.379'	0.423*
AD	0.001	n/a	0.001	n/a	0.001	n/a
L	- 0.000	n/a	- 0.000	n/a	- 0.000	n/a
AD/L ratio	n/a	0.014	n/a	0.103	n/a	0.243

*Significant codes: ' p-value<.1, * p-value <.05, ** p-value<.01, *** p-value<.001*

Clarke's tests showed that Model 2 fit significantly better across the anchor choices (p<.05, p<.01, and p<.001 for trimmed-mean-, median-, and NN-anchor respectively). However the coefficient of the distance-level ratio was insignificant. Thus, this variable was excluded and a re-analysis was done on the model below.

Model 3: $$\lambda(x_j^n) = \alpha + \beta \cdot CD(x_j^n) \quad (4)$$

Table 1.9 shows regression results of Model 3. The differences in the sizes of the remaining coefficient and the intercept between Models 2 and 3 were hardly discernible. Clarke's test significantly favored Model 3 over Model 2 (p<.001 with any anchor).

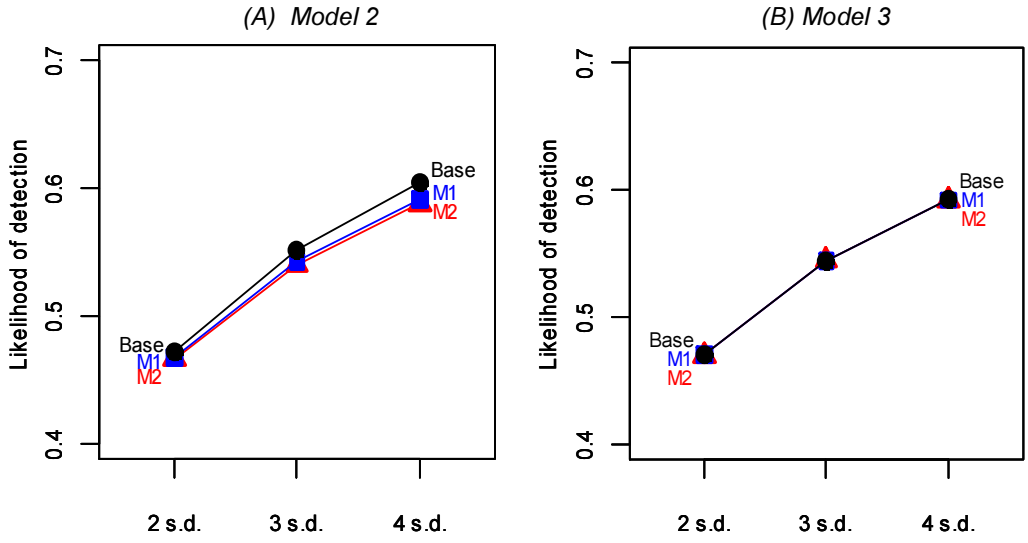
Table 1.9. *Logistic regressions of outlier identification likelihood (M-scale)*

	Anchor		
	Trimmed-mean	**Median**	**NN**
Intercept	- 2.261*	- 1.116*	- 0.713'
CD	1.347*	0.587*	0.447*

*Significant codes: ' p-value<.1, * p-value <.05, ** p-value<.01, *** p-value<.001*

I compared the fits of anchors using Model 3. By Clarke's tests, NN turned out to be significantly worse than the median (p<.05). It was also worse than the trimmed-mean but not significantly (p≈.34). The median fitted only slightly better than the trimmed-mean but not significantly (p≈.91). These results suggested that the median was the anchor. The plot (Figure 1.5) of the median predicted likelihoods of outlier identification by Models 2 and 3 using the median as an anchor shows no meaningful M-scale effects.

Figure 1.5. *Median predicted probabilities of outlier identification (M-scale)*



1.4.2.2 Outlier perception and forecast aggregation

Aggregate anchor. From this experiment also, SDMs centered around the median (Table 1.10), and the location tests (Wilcoxon tests) could not reject the median as the center of aggregation for all scaling and outlier levels.

Table 1.10. *Median standardized departures from the median (M-scale)*

	Base scale	**A1 scale**	**A2 scale**
2 s.d.	0.03	- 0.03	0.19
3 s.d.	0.06	0.04	0.06
4 s.d.	- 0.15	0.08	0.21
overall	0.00	0.01	0.12

The results from quantile regressions examining the relation between perception and aggregation schemes (Table 1.11) showed low and insignificant effects of outlier weights to amounts bet in all scaling manipulations, in fact even lower than results from the Base level in the A-scale experiment. Again, participants chose to be conservative by focusing on medians regardless of how the data had been scaled.

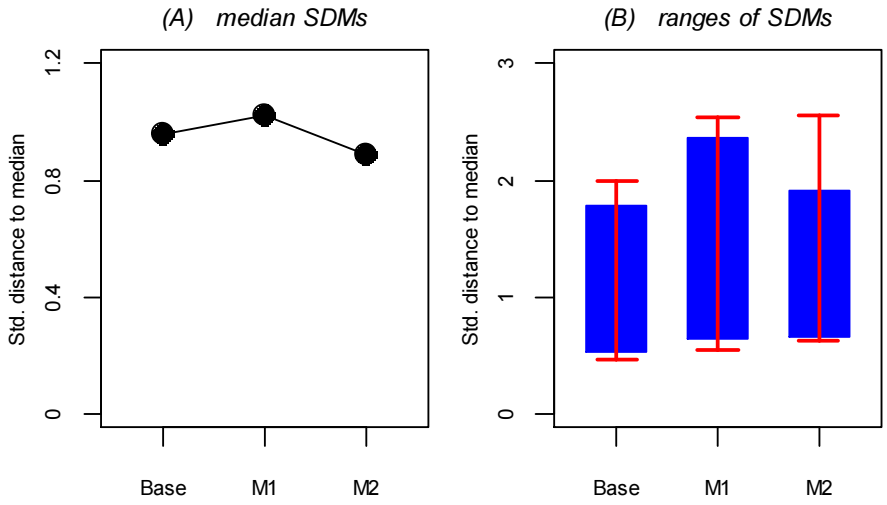
Table 1.11. *Median regressions of aggregation scheme on outlier perception (M-scale)*

	Base	**A1**	**A2**
Intercept	- 0.172	0.269	0.184
Bet	0.011	- 0.019	- 0.002

None of the coefficients were significant

Subjective variability and forecast aggregation. As shown in Figure 1.6, SDMs were not dissimilar among different scaling levels (Median=0.96, 1.02 and 0.88 for outlier levels of 2, 3 and 4 s.d. respectively).In fact none of the pair-wise Brunner-Menzel tests were significant. Similarly, in terms of dispersion of SDMs, all three scaling levels possessed comparable ranges, and Mood's test showed no significance in either of the scaling pairings.

Figure 1.6. *Medians and ranges of standardized departures from the median(M-scale)*



1.4.2.3 Discussion

Although participants in this experiment did exhibit differential placements in bets, that more extreme outliers received lower bets, the effect was not significant even at the Base level. So it was not entirely clear how multiplication scaling might have impacted how participants compared outliers among different degrees. However as among all scaling levels the median bets were exactly equal, and the distributions of bets largely covered a similar range, the results of this experiment suggested that multiplicative scaling did not impact participants' detection of outliers. As in the earlier experiment, participants here also anchored their aggregation around the medians of forecasts sets, regardless of how likely outliers seemed to them.

1.5 General discussion

People have a tendency to assign believability values to opinions, and these values are reflected in how opinions are combined. However in a one-shot situation, usually with no external information, they have to rely on their presumption of the natural distribution of opinions, often derived merely from the handful of opinions that happen to be available. However, this habit is beneficial only if their believability evaluations are fairly accurate.

This chapter focuses on DMs' treatment of outliers in aggregation, the issue that only a few studies have investigated. I started by examining how people judge a forecast

to be extreme. The results of two experiments suggested that participants judged the extremity of a forecast by comparing its distance from the others to the similarly judged distances of the other forecasts. This in fact closely follows the definition of an outlier. However this comparative distance index cannot guarantee the accurate identification of an outlier. Results of the simulation showed that the relation between the index and the true extremity of a forecast was unclear.

Moreover people's perspective on forecasts can be affected by mere scaling which is not an uncommon occurrence. As information is passed around, forecasts are scaled to suit the decision environments, interests, or preferences of the receivers. Operation directors might be more interested in overall performance of the business as they are concerned with total revenues, while sales directors might focus on year-on-year changes in their company's earnings. Managers of a Tokyo subsidiary make forecasts using information based in Japanese yen but will need to transform them into U.S. Dollars when reporting to headquarters in New York. An increase in the data level or A-scaling, as in the former example, has been found to attenuate subjective variability inference (which the examination of participants' aggregates in the current study also indirectly corroborated), so it is natural to assume that it will also affect outlier perception. I not only tested the impact of the data level on outlier identification, but also investigated how it affected the judgment process. Experimental results confirmed the effect of A-scaling where it was found that participants used a data level as the base of the ratio for the distance that a potential outlier lies away from the rest of the forecasts. However, the latter result anticipated that if the scaling was multiplicative, i.e. M-scaling, as in the yen-to-dollars example, there would be no impact. Indeed, results of the second experiment demonstrated no M-scaling effect.

But probably the more important question is how people use an outlier when aggregating forecasts. Unlike previous research, which assumed that participants correctly recognized outliers, this study innovates by examining the relation between subjective believability of an outlier and its weight in aggregation. In general, I found no correlation between the two; participants anchored their estimates around sample medians. However the results suggested that participants were more willing to include the planted outlier in their aggregates, albeit only slightly, if they perceived large forecast variability such as when the data are under a high level of A-scaling.

Despite the results of experiment in this study, we cannot conclude that take-a-median is DMs' definite invariant aggregation strategy. Participants might have felt that

they had too few forecasts to risk including the extreme one. But, when facing a potential outlier, take-the-median is still a risky strategy as outlier identification proves to be an error-prone task (see Appendix 1.1). Instead, even when an extreme forecast is present, and especially when a larger number of forecasts are at hand, unit weighting remains a prefer aggregation strategy, as much research has suggested(see also Appendix 1.2). Future research should explore aggregation strategies as a function of the number of forecasts, maybe controlling for subjective perceptions of variability. If the shift in an aggregation heuristic towards take-the-mean when there are more forecasts is observed, it is still not clear that DMs understand the benefit of unit weighting in conjunction with the value of additional forecasts. It will be interesting to see if DMs will actively acquire or are willing to pay for extra forecasts (and employ take-the-mean), or they rather resort to take-the-median as Sir Francis Galton would have recommended (Galton, 1907a; Galton, 1907b).

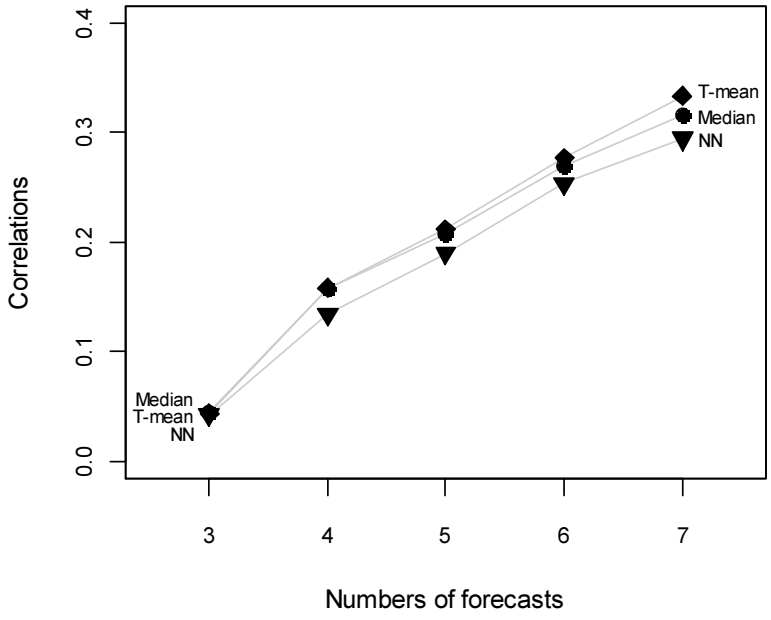
Appendix 1.1

To compare the 3 anchors (trimmed mean, median, and nearest neighbor) in term of their accuracies in evaluating an outlying degree of a forecast, I calculated the correlations between comparative distances based on each anchor choice as defined by Equation 1.2 in the section 1.2.1 of the chapter and the distances to the true mean, using the simulated forecast sets. In the case of 3 forecasts I simulated 10,000 sets of 3 numbers using a standard normal distribution ($\mu = 0$, $\sigma = 1$). To examine the case of 4 forecasts, I added 1 extra forecast simulated following the same distribution to each set produced for the 3-forecast case. Generally, the forecast sets used in the following case with *n* forecasts proceeded in the similar manner, i.e. by adding 1 extra forecast to the sets used in the previous case with *n-1* forecasts. Since the interest was in identification of the extreme forecast, only the most extreme forecast in each set was taken into calculation of the correlations. The most extreme forecast under the chosen anchor had to satisfy two criteria, 1) it must be the highest or the lowest forecast of its set, and 2) from the 2 forecasts satisfying the first criterion, it must be the one having the largest comparative distance based on that anchor.

The results are shown in Figure 1.7. We can see that each anchor choice did not yield substantially different performances. However, trimmed mean gained more

advantage as the more forecasts were added to the set, while median was its comparable alternative for an anchor especially if the forecast set was not too large.

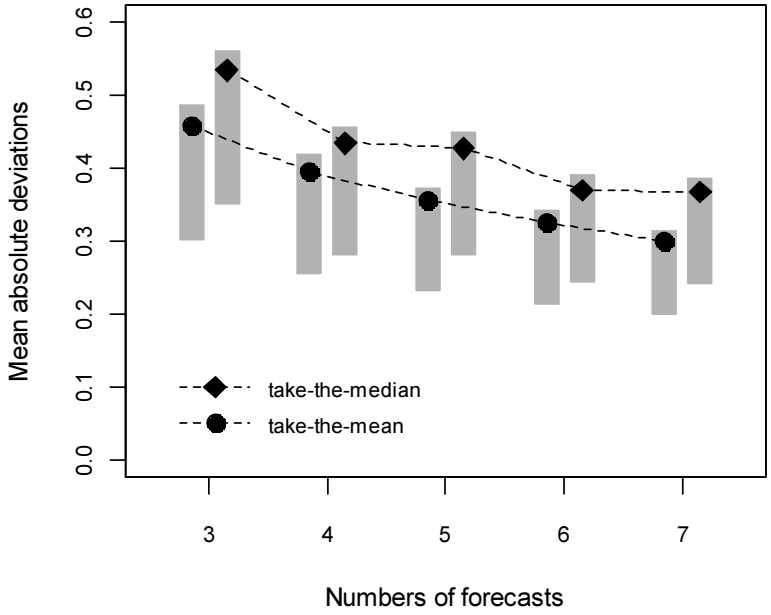
Figure 1.7. *Correlations between comparative distances and distances to the true mean*



Appendix 1.2

The forecast sets examined here were those simulated for Appendix 1.1. From each set, the take-the-mean and take-the-median aggregates could be directly derived..Figure 1.8 shows the mean absolute deviations (MADs) from the true mean (μ) by both aggregation strategies confirms that the taking the mean is on average more accurate than taking the median, but not by much. (Take-the-median results in MADs 9.7% - 22.0% larger than take-the-mean.). One interesting result here is that the accuracy of the take-the-median strategy using the even number of forecasts is not discernibly different from the accuracy of the same strategy using 1 few forecast. In fact, a take-the-median aggregate of the set with even-number forecasts partially includes a take-the-mean strategy since it is an average of the middle 2 forecasts. So assuming each forecast incurs a cost, and one has already obtained an even number of forecasts, it might be advisable not to do seek more forecasts if one's budget permits an acquisition of only one additional forecast.

Figure 1.8. *Mean absolute deviations from the true mean under take-the-mean and take-the-median aggregation strategies.*



Note: Grey bars cover the middle quintile of the deviations

2. Effects of outlier appearance order in aggregation of short forecast sequences

Abstract.

Experts' forecasts are often obtained serially. This has been demonstrated to affect people's ability to detect outliers, and consequently their choices of aggregation. Results from two experiments with short forecast sequences (length = 4) showed that aggregation processing modes also played a role. When participants produced and updated their estimates with each new forecast along a sequence, they missed an outlier if it appeared early and their final estimates showed no influence of that outlier; on the contrary, an outlier that appeared last got detected and ended up with a substantial weight in the final aggregation. When participants were made to recall earlier forecasts in a sequence and to produce only one single estimate at the end of each sequence, they duly realized forecast extremity, in spite of which outliers still exerted a weighty influence over the final aggregation regardless of order of appearance.

2.1 Introduction

When seeking opinions, one will not always summon several experts at once. Rather, one may only get to meet them separately and solicit their opinions one at a time. Such piecemeal nature of information acquisition could impact how a decision maker (DM) processes the inputs, and as a consequence results in a final decision different than if the same set of information were laid out in front of a DM in its entirety. Research has demonstrated that information obtained last appears to be more influential towards final judgments and decisions than those obtained first. This recency effect has been found in various fields such as finance (Reneau & Blanthorne, 2001; Guiral-Contreras et al., 2007), consumer choice (O' Brien & Ellsworth, 2012), medicine (Chapman et al., 1996), and law (Furnham, 1986; Kerstholt & Jackson, 1998; Dahl et al., 2009). In some situations, however, research has revealed the opposite effect, called primacy, where earlier information exerts a stronger impact (Carlson et al., 2006; Xu & Kim, 2008). Moreover, information used in earlier decisions has been found to influence later decisions despite its irrelevance (Curley et al., 1988; O'Reilly et

al., 2004). Even preference formation of consumers is subject to the order in which choices are presented (Matonakis et al., 2009; Carney & Banaji, 2012).Researchers have explored a variety of mechanisms and reasons why such order effects arise, including the direction that information is evaluated (Mantel & Kardes, 1999; de Bruin & Keren, 2003), the decrease of attention as DMs process information (Yates & Curley, 1986; Xu & Kim, 2008), range-frequency theory (O'Reilly et al., 2004), predecisional distortion (Russo et al., 1998; Carlson et al., 2006), initial information-processing goal (Kardes & Herr, 1990), complexity of the task characteristics (Marsh & Ahn, 2006), and how information is encoded (Hogarth & Einhorn, 1992).

While most research on order effects in decision-making has focused on evaluation and contingency judgment, order effects have also been encountered in forecast aggregation. Participants in experiments by Levin (1976) who were asked to compute means of nine-forecast sequences were affected by the nature of the aggregation task, i.e. whether participants were asked to respond with a single estimate at the end of each sequence, or whether they needed to revise their prior estimates every time they received a new forecast, as well as by their expectations about the sequences, i.e. the likelihood that the sequences could contain outliers. When it was made explicit that every sequence had an outlier, participants' estimates duly discarded outliers regardless of the aggregation task. In fact their responses were comparable to estimates of the control group in which nine forecasts were shown simultaneously. But when there was no explicit information about the possibility of an outlier, weights given to outliers were not discounted. In addition, there was a stronger discount on an outlier from participants who received an implicit hint that each sequence might contain an outlier compared to those who received no information. However, this disparity disappeared when the task demanded sequential estimate revisions rather than a single end-of-sequence estimate. That is, participants incorporated an outlier into their final estimates with weights that were based on how much they would expect to see an outlier. When the likelihood was short of certainty, and the task involved sequential updating of estimates, participants acted as if they disregarded information on outlier likelihood. In addition Levin (1976)also found that participants' responses were subject to a recency effect, especially when the task demanded sequential updates.

Building on the work of Levin (1976), the aim of this chapter is to explain the processes of sequential forecast aggregation and why an order effect might occur, particularly when a sequence contains an outlier. The framework I use is adapted from

Hogarth and Einhorn's (1992) belief adjustment model. I argue that when DMs are asked to evaluate comparative likelihood of forecasts and make only a single estimate at the end of a forecast sequence, all forecasts are likely to be recalled. This facilitates outlier detection, and consequently the final estimate will result from anchoring on the sequence's median with a small adjustment towards the outlier regardless of the order in which an outlier appears. However, when DMs initially produce estimates using the first few forecasts and then serially update their estimates as each new forecast is acquired, they tend to rely heavily on the compressed information in the form of the last revised estimates, rather than recall the full set of forecasts. This incomplete recollection results in giving an excessive weight to an outlier. So when an outlier appears last, its contribution to the final estimate ends up being significant.

I conducted two experiments to study these hypotheses, each with task environments designed to invoke the use of the different aggregation processes. In the first experiment participants gave and updated their estimates as each new forecast was presented. Their estimates fitted the sequential update process, and outlier aggregation weight displayed recency, that is, when outliers appeared last they would be incorporated into estimates substantially, resulting in estimates that deviated from medians significantly towards outliers. In the second experiment participants gave comparative accuracy ratings to forecasts, and made estimates only at the end of each sequence. Participants appeared to be cognizant of the extremity of forecasts, and their responses fitted a median-anchoring process. Estimates showed no significant order effect, but deviated significantly albeit slightly from medians.

2.2 Theoretical framework

In belief updating, DMs adjust their own perceived probability that a certain hypothesis is correct based on each new piece of evidence. In sequential forecast aggregation, DMs update their numerical estimates of a certain quantity using a new piece of forecast of the same quantity. With this general similarity, I see that the estimate-updating process can be modeled in the same fashion as the belief adjustment model (Hogarth & Einhorn, 1992).

2.2.1 Belief adjustment model

Hogarth and Einhorn (1992) propose that decision-makers update their belief in an anchoring-and-adjustment process, described generally as

$$S_k = A + w_k \cdot \left[s(x_k) - R \right] \tag{2.1}$$

where

x_k= k^{th} piece of evidence

S_k= degree of belief after evaluating k pieces of evidence

A = anchor, or prior belief before receiving the k^{th} piece of evidence (i.e. S_{k-1})

$s(x_k)$ = subjective evaluation of the k^{th} piece of evidence

R = reference point against which evidence is evaluated

The characteristics of the reference point, the anchor, and the weight function depend on how the information is processed, which can largely be put into two categories, the step-by-step (SbS) mode and the end-of-sequence (EoS) mode.

2.2.1.1 The SbS processing mode

This mode refers to when the current belief is adjusted, or re-estimated, each time the new piece of evidence x_k is obtained. The weight function depends on whether the new evidence is confirming or disconfirming compared to the reference point, that is whether the impact of the evidence, i.e. $\left[s(x_k) - R \right]$, is negative or positive, and the weight is proportional to the level of adjustment allowed on the direction of the evidence impact. The belief adjustment for this process can be expressed as

$$S_k = \begin{cases} S_{k-1} + \alpha S_{k-1} \cdot \left[s(x_k) - R \right] & when \quad s(x_k) \le R \\ S_{k-1} + \beta (1 - S_{k-1}) \cdot \left[s(x_k) - R \right] & when \quad s(x_k) > R \end{cases} \tag{2.2}$$

where

α =sensitivity towards negative evidence

β = sensitivity towards positive evidence

If DMs evaluate the new evidence against the current belief, i.e. $R = S_{k-1}$, the process will leads to moving-average like results, and recency is predicted when the information contains both confirming and disconfirming evidence.

2.2.1.2 The EoS processing mode

This mode refers to when DMs encode information by evaluating the impact of each piece of evidence in relation with other piece in the set, i.e. $\left[s(x_1,...,x_k) - R\right]$, to make a single judgment at the end of the evidence sequence. Thus the belief adjustment under this process can be described as

$$S_k = S_0 + w_k \cdot \left| \; s(x_1,...,x_k) - R \; \right| \tag{2.3}$$

With no prior belief S_0, DMs might just adopt the evaluation given to the first piece of evidence $s(x_1)$ as an the anchor yielding some degree of primacy.

Which mode of encoding is at work depends on the mode of response, and the characteristics of the evidence. If the task asks DMs to explicitly produce or update a judgment with every new piece of evidence, i.e. the SbS response mode, it is likely that the processing mode will be SbS in that that the task demands the estimation mode of information encoding. But if the task requires only a single final judgment after the sequence ends, i.e. the EoS response mode, EoS processing is not always guaranteed since the task does not lead DMs to any specific mode of information encoding. One noted difference between two encoding modes is the demand on memory and information-processing load. When the evidence is complex and/or the sequence is long, DMs may opt for the less demanding estimation encoding mode, hence they use the SbS processing mode even when the task only calls for the EoS response mode.

2.2.2 Estimate update model

How DMs give aggregation weights to forecasts could be based on their perception of the level of variability in the forecasts. The experimental result of Yaniv and Foster (1995) in which participants' use of a forecast depended on the level of its precision maybe an indirect support for this argument, as precision transmits the image of high confidence, and the accompanying narrow variance. So I assume that a forecast's aggregation weight is affected by DMs' perception of variability. I adapt the

belief adjustment model for a sequential forecast aggregation task by integrating the impact of subjective variability.

2.2.2.1 Weighting an outlier in the SbS processing mode

When forecasts F_n appear one at a time, it is likely that DMs initially produce estimates early on, and then update their estimates when a new forecast is obtained. These estimates are basically what DMs *believe* to be the correct value that forecasts are aiming for. As in the belief update model, DMs update their estimates by anchoring their judgments on the preceding estimate (i.e. current belief) and adjusting that estimate towards the latest forecast (i.e. new information) they have just received. The level of adjustment depends on how accurate that new forecast is perceived to be. This sequential updating process can be expressed by

$$E_n = E_{n-1} + u_n \cdot \left(F_n - E_{n-1}\right) \tag{2.4}$$

where

E_n = estimate updated with n^{th} forecast

u_n= update weight for F_n based on its perceived accuracy likelihood

Equation 2.4 can be rearranged into a form that implies that the new updated estimate is just a weighted average between the current estimate and the new piece of forecast, or $E_n = (1-u_n)\cdot E_{n-1} + u_n \cdot F_n$. Practically, aggregation under this process can also be thought of as weighted averaging of all forecasts in a sequence, or

$$E_N = \sum_{n=1}^{N} w_n \cdot F_n \tag{2.4a}$$

where the (final) aggregation weight w_n for each forecast can be calculated using its own update weight together with weights of other succeeding forecasts as

$$w_{N-i} = \begin{cases} u_{N-i} & ;i = 0 \\ u_{N-i} \cdot \prod_{k=0}^{i} (1-u_{N+1-i}) & ;i > 0 \end{cases} \tag{2.5}$$

In order to economize their memory load, it is likely that DMs will treat their current inference of variability along with their inference of mean (i.e. current estimates) as a summary of all forecasts acquired so far. However, to correctly update the two inferences, DMs would need to recall all previous forecasts. The implication of the incomplete recollection of forecasts is particularly important when a sequence contains an outlier. The result is outlier-recency, that is an excessive combination weight given to an outlier when it appears last in a sequence.

When an outlier appears early, before its extremity can be realized, it is obvious that DMs will fully incorporate an outlier in their estimates, and the sample from which DMs infer variability will include an outlier. As a sequence continues and a new forecast appears, if all previous forecasts are not recalled, an outlier will remain undetected. Consequently variability will not be attenuated and current estimates not corrected. However aggregation of succeeding regular forecasts will discount the original weight of an outlier, and eventually the influence of an outlier will fade away.

When an outlier appears last, DMs' ability to detect that outlier depends on how much of the sequence they can recall. If the recollection is complete, an outlier's extremity will be recognized in its full extent, and one can expect it to be appropriately discarded. If the recollection is short of perfect, DMs might be aware of an outlier's extremity only partially, and they will give a generous weight to an outlier when updating their estimates. Eventually the final estimates will feature a substantial aggregation footprint from an outlier.

The cause of the failure to correct inferences of a forecast sample might be more than just the incomplete recollection of the sample. DMs tend to uni-directionally evaluate information (Houston, Sherman & Baker, 1989; Mantel & Kardes, 1999), meaning a piece of information is judged based on information that precedes it, but a judgment already made is often not re-evaluated in light of the new information. So it is unlikely that an outlier already misjudged as regular will be re-evaluated. This implies that DMs will not correct or adjust their current estimates before combining the new forecast, even though the new forecast might indicate in retrospect how erroneous those estimates have been.

2.2.2.2 Weighting an outlier in the EoS processing mode

When DMs produce estimates only at the end of a sequence, there is no inference of a forecast set to rely on, so they will resort to recalling all forecasts

$F = \{F_1, F_2, ..., F_N\}$. The complete recollection is equivalent to a simultaneous forecast presentation, with all forecasts being shown together in DMs' mind rather than before DMs' eyes. And as a consequence it should provoke a similar aggregation process that the final estimate is a weighted average of forecasts around the sequence's (subjective) central value with a weighting scheme based on accuracy likelihood of each forecast, or

$$E_N = C + \sum_{i=1}^{N} w_i \cdot \left| \; F_i - C \; \right| \tag{2.6}$$

where

E_N = estimate after receiving Nforecasts

C = central value of the sequence

w_n = adjusting weight for F_n based on its accuracy likelihood, and $\sum_{i=1}^{N} w_i = 1$

The right side of Equation 2.4 can be reduced to a formula for a weighted average of all forecasts $E_N = \sum_{i=1}^{N} w_i \cdot F_i$, hence the adjusting weights are in practice equivalent to the aggregation weights.

However when DMs encounter an extreme forecast, they tend to focus their aggregates on the median (Yaniv, 1997; Harries et al., 2004; see also Chapter 1).So under EoS processing where full-sample recollection facilitates detection of an outlier (any distant forecast can be seen more clearly when contrasting with multiple regular forecasts), I expect DMs to arrive at the aggregate by anchoring on a sequence's median, and any departure from there towards an outlier should be only to an insignificant extent. In this sense, final aggregates can be expressed as

$$E_N = Md + d \cdot (Ol - Md) \tag{2.7}$$

or

$$E_N = (1 - d) \cdot Md + d \cdot Ol \tag{2.7a}$$

where d is a deviation weight, Md is the median of the sequence, and Ol is an outlier.

To summarize, when a task demands a sequential update of estimates, DMs are likely to under-discount an outlier, and give it an excessive update weight. However, the impact of that outlier becomes lower as more and more forecasts appear and get

combined into the newly revised estimate. On the other hand, when a task demands only a final estimate at the end of a forecast sequence and DMs evaluate each new forecast against those that come before, an outlier is likely to be detected and DMs' final estimates will be close to the medians of the sequences.

2.3 SbS-response experiment

This experiment required participants to explicitly produce intermediate estimates mid-sequence. However the design could not rule out the possibility that participants would try to recall the whole sequence when calculating a final estimate as the end of a forecast sequence.

2.3.1 Methods

Procedures. The experiment was conducted in a laboratory in Barcelona in Spanish on personal computers. Participants were given the instruction sheets which were also read out loud to them. The instructions stated that the experiment was about sales forecasting of one supermarket chain which operated stores across the southeastern United States. This supermarket held annual meetings where senior managers and executives from different departments gathered. At these meetings, the executives gave the forecasts for the sales of the following year of the stores under their supervision. The instructions further specified that the data came from the 2002 annual meeting hence the real sales were known. Participants were informed that the experiment contained 18 rounds concerning sales of 18 different stores, one store per round. In each round they would be given a set of four forecasts of monthly sales in units of thousand U.S. dollars of one particular store randomly selected from a pool of forecasts given at the company's annual meeting.

In the first stage of each round two forecasts, F_1 and F_2, appeared on screen (F_1 was positioned above F_2). Participants then were asked to give their own monthly-sales estimates of that supermarket based on the information they had received, and they would be rewarded according to the accuracies of their estimates. Estimates were allowed up to one decimal. Next, in the second stage where the third forecast, F_3, appeared on screen, participants were asked to revise their previous estimates based on all forecasts they had seen in any way they liked. They would also be rewarded

according to the accuracies of the revised estimates. In the third and final stage where the fourth forecasts, F_4, appeared, participants were again asked to revise their own estimates based on four forecasts any way they liked. Rewards would be given according to the accuracies of the newly revised estimates. The reward scheme in all three stages was that, an estimate that deviated from the realized (i.e. true) sales figure below 1 unit would earn 120 points, an estimate that deviated at least 1 but below 3 units would earn 60 points, and an estimate that deviated at least 3 but below 10 units would earn 20 points. Deviations of 10 or more units earned no points. With this payoff structure, the best strategy was to give the most accurate estimate possible. The exchange rate for payoffs was 50 points for 1 Euro.

Stimuli. Forecasts were chosen to follow a normal distribution with a mean of 85 and a standard deviation of 17. In each round, a participant encountered a sequence of four forecasts in one of the 18 outlier conditions. An outlier was either of 2, 3, or 4 standard deviations (s.d.), was either to the left or to the right of the population mean, and appeared either as the second forecast F_2, third forecast F_3, or fourth forecast F_4 (let's call these appearance orders as OF2, OF3, and OF4 respectively). Non-outlier forecasts were randomly generated and rounded to the nearest integer. They were restricted to lie within 2 s.d. from the mean to avoid being extreme.

Table 2.1. *Rotation of non-outlier forecast sets and conditions for each participant.*

	Participant 1	**Participant 2**	**...**	**Participant 19**	**Participant 20**
Condition 1	set 1	set 2	...	set 19	set 20
Condition 2	set 2	set 3	...	set 20	set 1
Condition 3	set 3	set 4	...	set 1	set 2
:	:	:	:	:	:
Condition 18	set 18	set 19	...	set 16	set 17

Only twenty sets of three non-outlier forecasts were simulated and used in rotation among all participants and all conditions (Table 2.1), and the three forecasts in

each set appeared in a fixed order of comparative appearance.[5] Since for any single participant each outlier would appear three times, the forecast sets needed to be level-adjusted to avoid the detection of the stimuli pattern by adding a fixed integer to all forecasts in each sequence. The level-adjusting integer for each sequence was randomly and independently selected from a uniform distribution ranging from -3 to +3. In which round experimental participant encountered each condition was randomly selected individually and independently.

Participants. Twenty participants aged between 18 and 22 years took part in this experiment. Twelve were female, eight were male. They were recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the experimental laboratory. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 10.11 Euros.

2.3.2 Results

2.3.2.1 Aggregation process

To test that the task of the experiment prodded participants to encode information and then process the final aggregation as designed or not, I compared the fit of estimates to each process's model. The comparison was made for each outlier appearance order separately, which allowed the tests to focus on the fittings of both processes independent of the outlier orders. Also, since different outlier degrees could affect differentially variability inference which in turn influenced how a new forecast would be incorporated during an update, average combining weights were not restricted to be the same across all outlier degrees. As the main concern was about how participants came to their final estimates, the analyses focused on final estimates, E_4. For the SbS process the regressions were based on Equation 2.4 involving updating intermediate estimates E_3 to final estimates E_4 using the fourth forecasts F_4. For the EoS process, regressions were based on Equation 2.7a involving the values of final estimates E_4 in term of deviations towards outliers from medians. Taking into account that each participant gave multiple responses, I used linear mixed models to capture

[5] Call the three non-outlying forecasts in a sequence X_1, X_2, and X_3. Under OF2 condition, the order of appearance would be $\{X_1, OI, X_2, X_3\}$; under OF3 condition, $\{X_1, X_2, OI, X_3\}$; and under OF4 condition, $\{X_1, X_2, X_3, OI\}$.

within-subject correlations.[6] And since the two linear mixed models were non-nested, Clarke's test (Clarke, 2003 and 2007) became the method of choice for model selections. Samples whose responses lay outside the range of forecasts were excluded from the tests.[7] Results are in Table 2.2. Although in all orders of outlier appearance the test results favored the SbS process, only when outliers appeared early on that the result was significant. It was possible that the late appearance of an outlier prompted some participants to question the regularity of a sequence, and consequently to revisit the whole sequence when calculating the final estimates. In this case, one could say that outliers that appeared early on remained undetected even at the end of their sequences, so participants continued with the SbS process to derive final estimates.

Table 2.2. *Comparing fits of models under SbS and EoS processes, SbS experiment*

	MAE from regressions		**Clark's model selections**
	SbS	**EoS**	
OF2	6.16	7.57	SbS*
OF3	6.62	5.85	SbS
OF4	10.00	9.81	indifferent

*Significant codes: * p-value <.05*

2.3.2.2 Forecast combination weights

Update weights.[8] From Figure 2.1 Panel A, we can see that outliers that appeared later received lower update weights than those that appeared earlier. This at

[6] Since all responses from experiments in this chapter was affected by within-subject correlations, all analyses were based on linear mixed models.

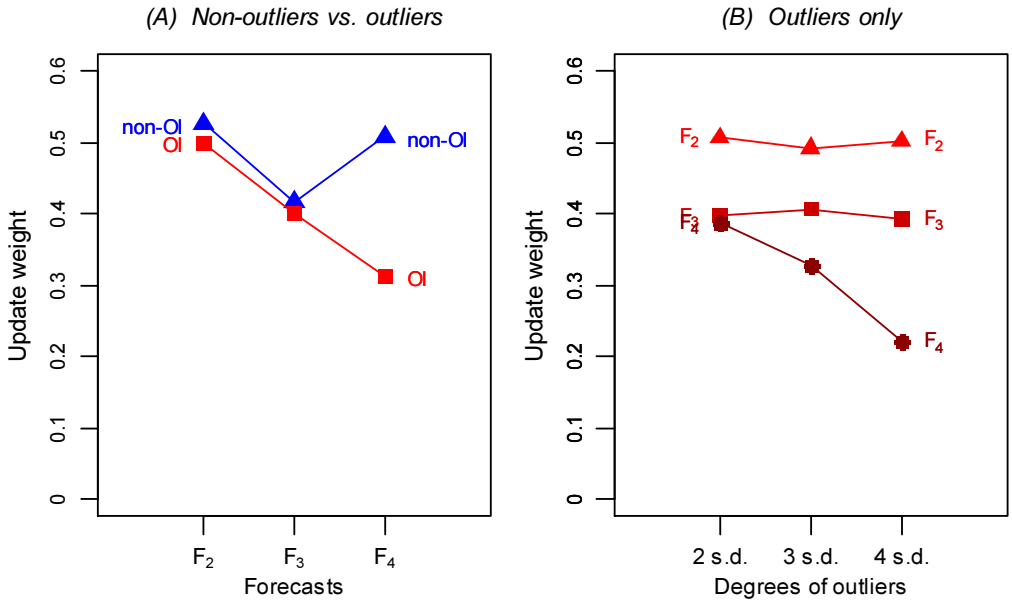
[7] At the first stage it was clear that estimates that lay outside the range of the two forecasts should be discarded. However at the later stages, estimates outside the range of the forecasts shown so far could be due to imperfect memory rather than merely be input mistakes. But to be on a conservative side, I decided to consider them irregular and to exclude them as well. Moreover, since the calculation of update weights depended on estimates in the previous stage, any analysis specific to one stage would ignore sample whose responses from one earlier stage were deemed irregular. That is, an analysis of responses from the sencond stage would exclude samples considered irregular in the first and second stages; an analysis of responses from the third stage would exclude samples considered irregular in the second and third stages. A total of 10, 16, and 22 samples were discarded in this analysis.

[8] Due to the formula for calculating update weights, samples whose intermediate aggregates were equal to the new forecasts to be combined in were discarded. There were a total of 4 of such samples, all of which happened with E_2 and F_3 in the OF2 condition.

first might look like a sign of participants' becoming more aware of outliers' extremity when there were more non-outlying forecasts that came before. If this was true, the more extreme forecast should have received a lower update weights, as a high level of extremity is associated with low accuracy likelihood. But Panel B suggests that such was the case only among F_4 outliers, i.e. when outliers came last ($F(1,79)=14.57$, $p<.0001$).

For F_2 outliers, the lack of a linear effect from outlier degrees on received weights was trivial, as detection was impossible in a sample of two forecasts. Hence we can see weights between outlying F_2 and non-outlying counterparts did not differ significantly ($M=0.500$ vs. 0.527, $F(1,284)=1.49$, $p\approx.22$). We can also see the lack of such a linear effect on update weights given to F_3. This could mean that participants did not detect those F_3 outliers either. The fact that F_3 outliers received less weight than F_2 outliers ($M=0.407$ vs. 0.500, $F(1,120)=6.08$, $p<.05$) outliers might be merely due to the sample size effect, i.e. forecasts from a large sample should individually contribute on average less to the aggregate than forecasts from a small sample should. As we can see, a similar pattern existed between F_2 and F_3 non-outliers as well.

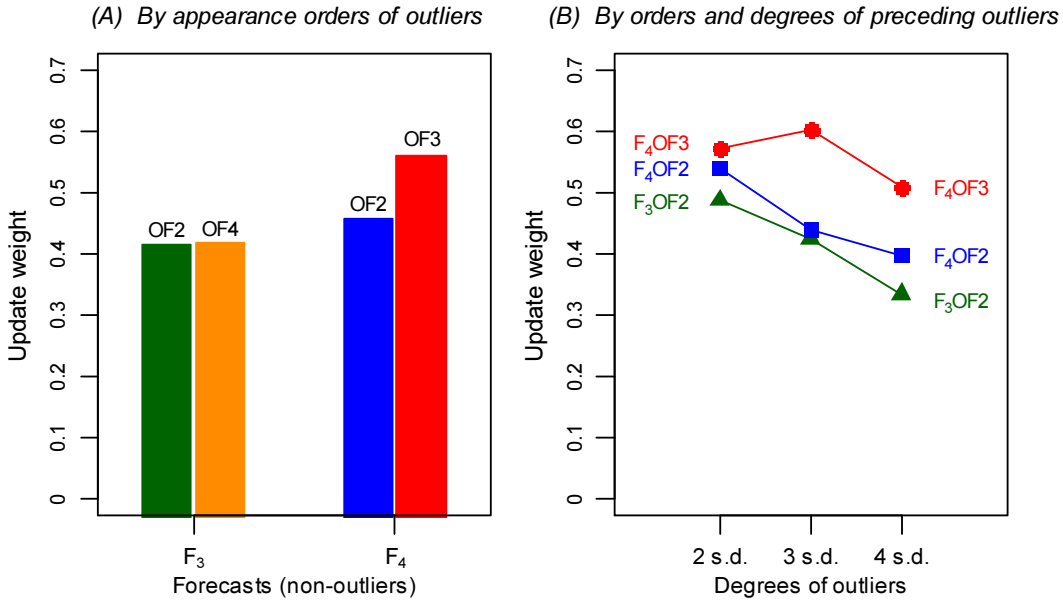
Figure 2.1. *Mean update weights given to new forecasts of different appearance orders*



The order that outliers appeared could also affect how non-outliers were incorporated into updated estimates. DMs could try to compensate for high update weights given erroneously to outliers earlier by giving even higher update weights to succeeding non-outliers. Experimental participants indeed gave higher update weights to non-outlying F_4 than to non-outlying F_3 (M=0.509 vs. 0.416, F(1,387)=1.63, p≈.20)[9] contrary to what one would expect due to sample size.

Figure 2.2 shows how update weights of *non*-outliers were impacted by the appearance orders of outliers. As shown in Panel A, the update weights U_3 for non-outliers F_3 were not dissimilar between under the OF2 (when they were preceded by an outlier) and under the OF4 (when they were not preceded by an outlier) conditions (M=0.414 vs. 0.418, F(1,174)=0.00, p≈.95). That is the levels of extremity of F_2 update did not influence the update weights given to F_3. This was expected as the analyses shown earlier suggested that an outlier was not detected when a sample contained only three forecasts

Figure 2.2. *Impacts of outlier appearance order on update weights of non-outliers*



[9] As the experiment had no data about how DMs would weight non-outlying F_4 when they were not preceded with an outlier, this comparison was only a proxy, especially if one assumes the sample size effect, which might be the reason that the difference did not reach significance. If a sample size four forecasts was expected to reduce update weights by 0.026 from weights given to a sample of three, the compensating raise of update weight of the fourth forecasts here would approach significance (p<.10); if the expected reduction was 0.048, the raise would be significant (p<.05).

As shown in Panel B, U_3 for non-outliers F_3 under the OF2 condition declined with along the increase of outlier degrees, that is the more extreme the outliers were, the less likely a succeeding forecast on average seemed to be. When a sample is expected to have wide variability the same *spot* forecast is less likely to be the true value than when the same sample is expected to have narrow variability as there are more spot forecasts that share a total likelihood to be accurate.[10] So these results suggest that F_2 outliers were not discarded when participants made variability inference. However, this effect was not significant ($F(1,179)=1.67$, $p\approx.20$).

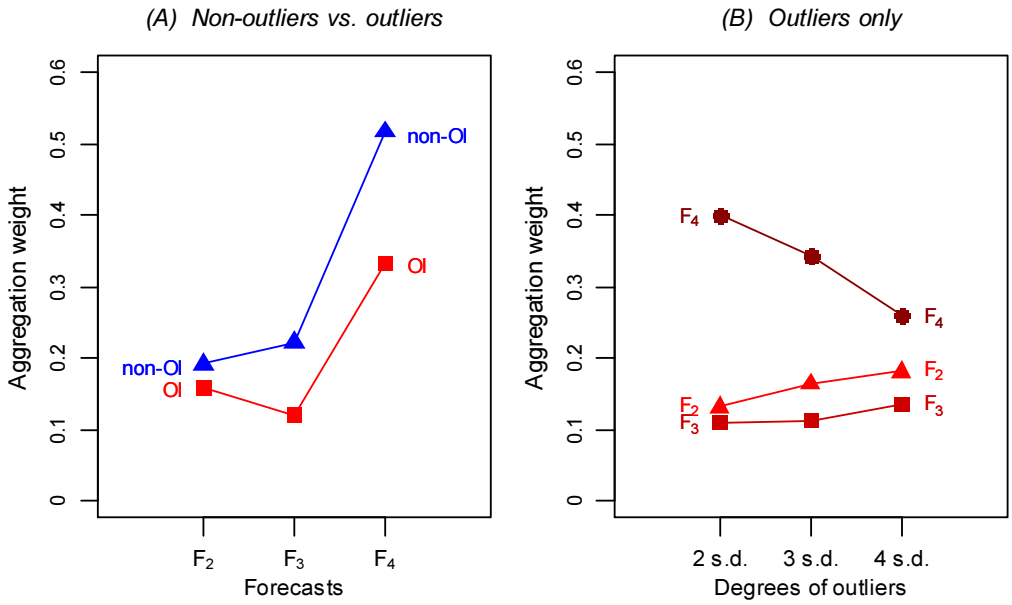
But with a sample size of four, update weights for non-outliers F_4 were discernibly impacted by the timing of outliers. As seen in Panel A for Figure 2.2, non-outliers F_4 under the OF3 condition, i.e. that were immediately preceded by an outlier, received higher weights than those under the OF2 condition, i.e. that were one stage removed from an outlier. However the difference was not significant ($M=0.562$ vs. 0.460, $F(1,172)=0.77$, $p\approx.38$). If participants indeed tried to correct high update weights for non-outliers, the urge to do so became weaker when preceding outliers were further in the past. In fact, declining update weights U_4 of non-outliers F_4 under the OF2 condition as shown in Panel B of Figure 2.2 suggested that outliers that appeared too early were not fully recognized and not entirely excluded from variability inference for the same reason that U_3 for non-outliers F_3 declined under the OF2 condition. But also this effect of outlier degrees was not significant ($F(1,179)=1.67$, $p\approx.20$). There was no similar trend with weights U_4 of non-outliers under the OF3 condition that just saw an outlier recently.

Aggregation weights. As we can see in Figure 2.3, the SbS aggregation task resulted in recency. Forecasts that appeared last in the sequence were aggregated with significantly higher weights than those that appeared earlier whether among non-outliers ($M=0.510$ vs. 0.206, $F(1,549)=30.35$, $p<.001$) or among outliers ($M=0.314$ vs. 0.140, $F(1,264)=30.79$, $p<.001$). Even though participants gave lower update weights to F_4 outliers, the impacts they had over final estimates were still stronger than those of outliers that appeared earlier as F_2 and F_3. This is significant when outlying degrees

[10] Consider sample A whose range that contains a true mean with 80% confidence level is from 35 to 40, and sample B whose range that contains a true mean with 80% confidence level is from 5 to 70. The likelihood that, for example, 35, 36, or 37 is a true mean of a sample is higher in sample A than in sample B.

were 2 s.d. (M=0.376 vs. 0.121, F(1,272)=19.42, p<.001), and 3 s.d. (M=0.319 vs. 0.135, F(1,77)=15.19, p<.001). Only the most extreme 4-s.d.outliers that participants discounted enough in the final updates that the eventual impacts did not differ significantly (M=0.240 vs. 0.162, F(1,76)=2.09, p≈.15). Notably, in the opposite direction of F_4 outliers, when appearing as F_2 or F_3, outliers of higher degrees ended up with higher aggregation weights. As explained earlier, this was due to the fact that outliers of those appearance orders were not discarded when making variability inference, so succeeding forecasts, as individual spot forecasts, seemed less likely when outliers were more extreme. Consequently they received lower update weights that would not eventually discount less aggregation weights of outliers. But the trends were not significant (F(1,79)=0.56, p≈.46 for F_2 outliers; F(1,73)=0.34, p≈.56 for F_3 outliers).

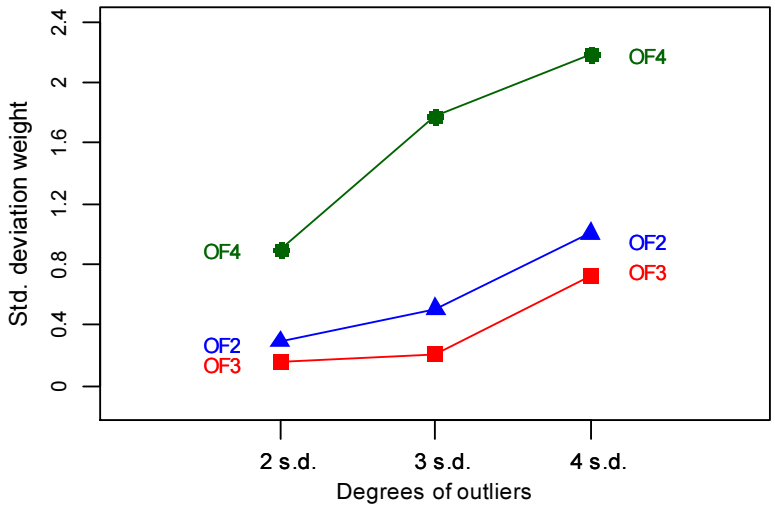
Figure 2.3. *Aggregation weights that forecasts had in final estimates, SbS experiment*



To compare the aggregation under this SbS process to a 'normative' scenario of simultaneous forecast presentation, I examined the deviations that final estimates strayed away from median towards outliers standardized by variability, since variability was hypothesized to affect aggregation. I chose to calculate sample variability by not including outliers, so the deviations were measured in the unit of mean deviation of

non-outliers from medians of their respective samples (similar to SDM in Chapter 1), so the deviations could be directly compared to the normative case that discards outliers. Also, unlike the deviation weights that are calculated straightforwardly as in Equation 2.7 which were practically in the unit of a distance between an outlier and a median, the unit of deviation measurement in this type of standardization allows comparisons across different levels of outliers' extremity. This measurement can be regarded as a standardized deviation weight (SDW). SDWs from final estimates were shown in Figure 2.4.

Figure 2.4. *Deviations of final estimates from medians towards outliers, SbS experiment*



Even though more extreme F_4 outliers received lesser update weights, the attenuation was so slight compared to their sizes that their footprints remained, so much that higher-degree F_4 outliers resulted in larger deviations than lower-degree counterparts ($F(1,84)=3.12$, $p<.10$). Eventually, at all outlier degrees, final estimates deviated significantly from medians (for 2.s.d. $F(1,19)=13.51$, $p<.01$; for 3 s.d. $F(1,18)=7.96$, $p<.05$; for 4 s.d. $F(1,19)=5.00$, $p<.05$).

Extreme forecasts that appeared as F_2 and F_3 exerted similar outlier degree effects ($F(1,91)=4.04$, $p<.05$ for F_2 outliers; $F(1,91)=4.04$, $p<.05$ for F_3 outliers), but the mechanics was different from F_4 outliers. For F_2 and F_3 it was through their influence on variability inference that, as explained earlier, more extreme outliers were

diluted less. At the end, under the OF2 condition only 2-s.d. outliers did not induce significant deviations ($F(1,19)=1.80$, $p\approx.20$), unlike 3-s.d. ($F(1,20)=5.61$, $p<.05$), and 4-s.d. outliers ($F(1,20)=6.53$, $p<.05$). While under the OF3 condition, both 2-s.d.($F(1,19)=0.49$, $p\approx.49$), and 3-s.d. ($F(1,19)=2.44$, $p\approx.14$) outliers did not result in significant deviations; deviations from 4-s.d. outliers only approached significance ($F(1,20)=3.45$, $p<.10$). That F_3 outliers did not cause significant deviations might be due to, as mentioned earlier, corrective weight compensation given to non-outliers that followed them.

Overall, in terms of deviation from medians towards outliers, we can also see outlier-recency, that is, when outliers appeared last the standardized deviation weights (SDWs) were much higher than otherwise ($M=1.22$ vs. 0.36, $F(1,306)=17.45$, $p<.001$).

2.3.2.3 Discussion

In this experiment which involved making an initial estimate and revising it each time a new forecast appeared, invoking the step-by-step information encoding, participants seemed to process information in the intended step-by-step fashion, although not significantly so in all cases. At the end of forecast sequences, estimates from this aggregation process exhibited outlier-recency. Despite the fact that participants appeared to recognize outliers when they showed up last as the fourth forecasts and discounted their weights accordingly, this discount was not sufficient thereby resulting in final estimates that deviated significantly from medians towards outliers.

On the other hand, when outliers appeared as the third forecasts, participants did not seem to recognize their extremity at the time of their appearance, and as a result those outliers were mistakenly given large update weights. But by the end of the sequences, i.e. samples reached the size of four forecasts, participants seemed to realize the mistake and tried to compensate for overweighting of earlier outliers by raising the update weights of the following fourth non-outlying forecasts in order to dilute aggregation weights of outliers in final estimates. This neutralized the influence of outliers, from which we can see that participants' final estimates did not deviate significantly away from medians.

However, outliers that appeared first as the second forecasts, which expectedly received weights equal to non-outlying first forecasts, remained undetected until the end

of sequences when generally the sample size was sufficiently large to facilitate outlier detection. Participants did not exhibit any attempt to dilute weights given earlier to outliers, suggesting that participants' retrospective detection reached back only one stage. While aggregation weights of these outliers ended up low thanks to dilution from update weights of succeeding forecasts, final estimates in many cases remained significantly distant from medians.

In the next experiment I will explore if outlier-recency will disappear when DMs are tasked to make only a single estimate at the end of each sequence.

2.4 EoS-response experiment

This experiment was designed to encourage a recollection of earlier forecasts in a sequence and avoid intermediate (mid-sequence) estimation by having participants evaluate each forecast when it appeared in comparison to other forecasts and asking them to give estimates only at the end of each sequence. However, the design could not rule out the possibility that participants would produce intermediate estimates which their final estimates would rely on.

2.4.1 Methods

Procedures. The same as in the SBS-response experiment. The only difference was the tasks that participants had to perform.

In the first stage of each round, only two forecasts (F_1 and F_2) appeared on screen. When experimental participants were ready, they moved on to the next stage where the third forecast (F_3) appeared. In this stage participants would be asked to rate the likelihood that this third forecast was the most accurate, i.e. closest to the real sales figure of that supermarket compared to the first two forecasts (F_1 and F_2) on 7-level Likert scale, from "0" being not at all likely to "6" being most likely. At their own pace, they then moved to the third and last stage where the fourth forecast (F_4) appeared to which they would be asked to give the accuracy likelihood rating in comparison to the first three forecasts (F_1 , F_2 ,and F_3).

In this last stage, experimental participants were also asked to give their own estimate regarding the sales of that supermarket which they could base it on the four forecasts in any way they liked. Participants were allowed to answer up to the first

decimal. They would receive rewards depending on the accuracy of their estimates. The reward scheme was the same as in the SbS experiment. The exchange rate for payoffs was 150 points for 1 Euro.

Stimuli. This experiment used the same set of simulated forecasts as the SbS experiment.

Participants. Twenty participants aged between 18 and 22 years took part in this experiment. Twelve were female, 8 were male. They were recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the experimental laboratory. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 10.29 Euros.

2.4.2 Results

2.4.2.1 Outlier perception

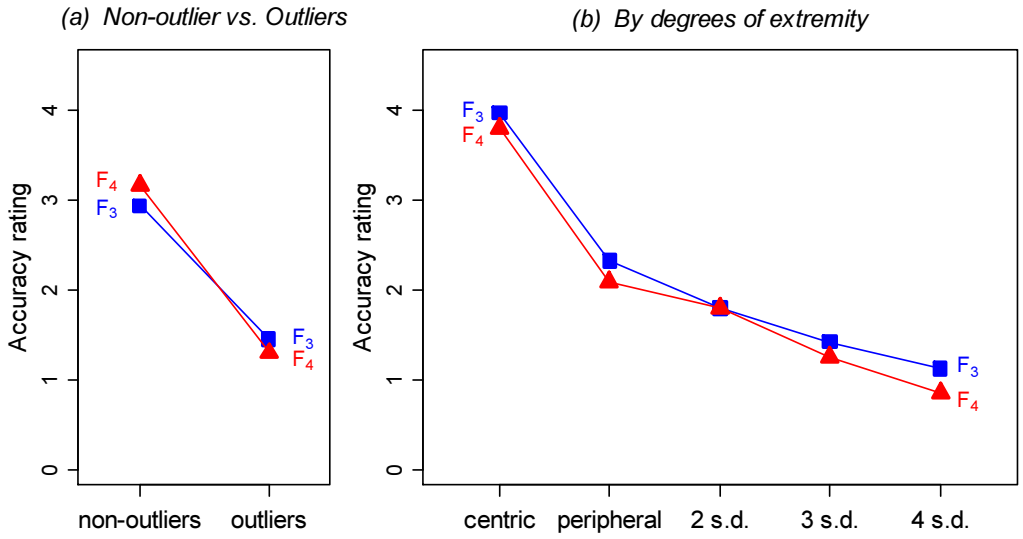
Generally, the more extreme a forecast is considered to be, DMs will perceive it as less accurate. Hence we could test if the experimental participants were aware of outliers by examining ratings they gave to forecasts. As shown in Figure 2.5 Panel A, non-outliers received significantly higher ratings than outliers[11], whether among the third forecasts F_3 (M=2.94 versus 1.45, F(1,339)=88.68, p<.001), or among the fourth forecasts F_4 (M=3.17 versus 1.30, F(1,339)=130.52, p<.001). And as Panel B depicts, even within the non-outliers, those that lay closer to the center of the sequence, or *centric* forecasts, received higher ratings than those that lay at the boundary of the sequence, or *peripheral* forecasts[12], whether among F_3 (M=3.98 versus 2.33, F(1,219)=84.75, p<.001), or among F_4 (M=3.81 versus 2.09, F(1,219)=89.64, p<.001). We can also see significant linear declines in ratings from the peripherals to outliers of 2, 3, and 4 standard deviations, both among F_3 (F(1,250)=41.80, p<.001), and among F_4 (F(1,188)=37.27, p<.001). Moreover, the ratings for F_3 tracked the ratings for F_4

[11] Taking into account that each participant contributed more than one responses, linear mixed models were used through out this study.

[12] A centric forecast is a forecast that is neither the highest nor the lowest among all forecasts having appeared so far; a peripheral forecast is a forecast that is either the highest or the lowest among all forecasts having appeared so far. For example, let the non-outlying forecasts be {1, 5, 13} and the planted outlier is {20}. If the appearance order is {1, 5, 13, 20}, at the second stage where a participant has seen only the first three forecasts, {5} is centric, while {1} and {13} are peripheral. However at the third stage where a participant has seen all forecasts, {5} remains a centric forecast, and {1} remains peripheral, but {13} is now centric.

closely, being only slightly lower but not to a significant degree (M=2.44 vs. 2.55, F(1,219)=1.16, p≈.28). These results demonstrated that participants were cognizant of forecasts' comparative extremity from the second stage when the sample size was three.

Figure 2.5. *Mean accuracy ratings given to forecasts, EoS experiment*



2.4.2.2 Aggregation process

Since participants had exhibited appropriate awareness of outliers' extremity since the second stage, which supported the possibility of sequence recollection, one could expect the EoS rather than SbS, process to be at work. To check this, I compared the fit of estimates predicted by each process's model. The comparison was made for each outlier appearance order separately so the fittings were independent of the outlier orders. As in the analyses of the SbS experiment, I did not restrict that on average extreme forecasts had the same combination weights across different outlier degrees. Equation 2.7a was used for the EoS process; and since participants in this experiment did not give intermediate aggregates, even if they ever calculated ones, Equation 2.4a became a proxy for the SbS process. Here I assumed that the first two forecasts were combined to make initial estimates with equal weights of 0.5 since they were presented together, which the results in the SbS experiment could corroborate. Again Clarke's test (Clarke, 2003 and 2007) was chosen for model selections. I also excluded samples

whose aggregates lay outside the range that forecasts from their respective sequences covered.[13] Results are in Table 2.3.

Table 2.3. *Comparing fits of models under SbS and EoS processes, EoS experiment*

	MAE from regressions		Clark's model selections
	EoS	**SbS**	
OF2	5.42	5.02	EoS **
OF3	5.71	5.65	EoS*
OF4	6.28	6.29	EoS **

*Significant codes: * p-value <.01, ** p-value <.001*

Overall both models gave a similar fit by the account of mean absolute errors (MAE), however in general SbS was likely to produce lower MAE for it contained more regressors. But the results of Clarke's tests, which feature the correction of such bias in the same fashion as the Schwartz information criterion, significantly favored the EoS process in all outlier appearance orders.

2.4.2.3 Forecast combination weights

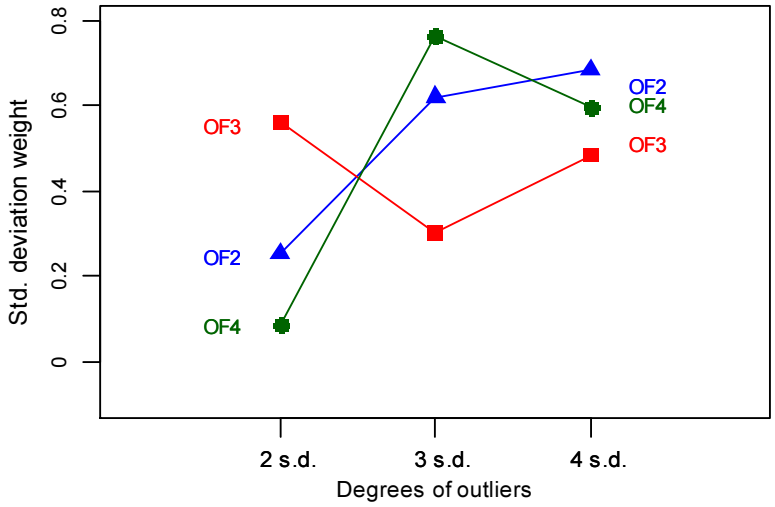
The results above strongly pointed towards EoS as the participant's operating process. Thus one would not expect to find an order effect, and that final estimates from participants in this experiment would anchor around sequences' medians and would not deviate far from there in any of the appearance orders. Figure 2.6 depicts deviations from medians towards outliers that final estimates exhibited, in a standardized form (SDW) as explained in analyses of the SbS experiment.

While under the OF2 condition the deviations showed some linearity along outlier degrees, the effect was not significant ($F(1,97)=2.43$, $p\approx.12$). No similar trend was found in the other conditions. In fact, there was no significant difference in deviations among different degrees of outlier whether under the OF3 condition ($F(1,96)=0.23$, $p\approx.79$), and the difference only approached significance under the OF4 condition ($F(1,94)=2.90$, $p<.10$). In most cases final estimates deviated significantly

[13] There were 2, 2, and 4 of such samples in OF2, OF3, and OF4 conditions respectively from a total of 120 samples in each condition.

away from medians. Under the OF2 condition, while deviations of 2-s.d. outliers only approached significance ($F(1,20)=3.74$, $p<.10$), those of 3-s.d. outliers ($F(1,20)=4.56$, $p<.05$), and 4-s.d. outliers ($F(1,20)=10.06$, $p<.01$) were significant. Under the OF3 condition, only 2-s.d. outliers did not depart significantly from medians ($F(1,20)=1.71$, $p\approx.21$), unlike 3-s.d. outliers ($F(1,20)=5.30$, $p<.05$), and 4-s.d. outliers ($F(1,20)=4.61$, $p<.05$). Similarly, under the OF4 condition, 2-s.d. outliers' deviations were not significant ($F(1,20)=0.76$, $p\approx.39$), while 3-s.d. outliers ($F(1,20)=6.99$, $p<.05$), and 4-s.d. outliers ($F(1,20)=7.70$, $p<.05$) deviated significantly from medians. Overall, the deviations were equal across all appearance orders with no order effect ($M=0.68$, 0.45, and 0.48 under the OF2, OF3, and OF4 respectively, $F(1,330)=0.07$, $p\approx.94$).

Figure 2.6. *Deviations of final estimates from medians towards outliers, EoS experiment*



2.4.2.4 Discussion

This experiment required participants to evaluate a forecast against other forecasts in its sequence that appeared earlier without having to explicitly produce intermediate estimates, to discourage participants engaging in step-by-step forecast updates, and to encourage recollection of forecast sequences. With this design participants appeared to be able to judge appropriately extremity of forecasts, even before a sequence ended when a sample size was three. From participants' responses, they seemed to arrive at their final estimates by the intended EoS process.

Expectedly, final estimates exhibited neither recency nor primacy. However, in the majority of cases, final estimates still deviated away from medians towards outliers to a significant degree, especially when outliers were more distant. So aggregates appeared as if participants had substantially included outliers despite the fact that they recognized the extremities. However the deviations were of the same level regardless of the outliers' levels of extremity.

2.5 General discussion

Normally when people search for forecasts, they will initially consult a few experts, and will continue to seek more opinions if time and other resources permit. People do not know with certainty how many forecasts they will finally obtain, or whether they will encounter an additional forecast before making a final decision. With such serial and indefinite nature of acquisition, it is natural to summarize forecasts into a decision-ready single estimate that can be updated later when additional forecasts become available. And in this situation, the timing that forecasts, especially outliers, appear can impact how they contribute to the final estimate crucial to decision-making. From the results of this study's first experiment, the step-by-step estimation-and-update mode can obscure an outlier from being detected especially when it appears early. Fortunately, as DMs obtain more forecasts, the significance of this outlier gets diluted further, and eventually aggregation will appear as if an outlier had been discarded. The problem, however, arises when an outlier appears last. In this case, despite detection, an outlier remains a weighty forecast within the final estimate.

As recency of an outlier's impact seems to stem from sequential updates of forecasts, when DMs instead aggregate all forecasts only at the end of a sequence, the timing when an outlier appears becomes irrelevant, which the results of the second experiment corroborated. Even though in this processing mode DMs can judge forecasts' extremity appropriately, the outlier detection is largely ineffective as it is not translated into a sufficient exclusion of outliers from aggregation. In this experiment, final estimates deviated significantly, albeit slightly, from the medians.

As we can see, each processing mode has its own advantage and disadvantage. The end-of-sequence mode obviates recency but is mentally taxing. The mental-load economizing step-by-step estimation-and-update mode is sufficient unless the search for information ends with an outlier; although that situation is statistically, and behaviorally

unlikely. In this study, the sequences of forecasts had fixed length, while in the real world DMs can continue to search for more forecasts. When evidence accumulation stops may depend on costs and benefit of extra information (Gilliland, Schmitt & Woo, 1993; Hulland & Kleinmuntz, 1994; Saad & Russo, 1996), on time pressure (Hulland & Kleinmuntz, 1994), as well as whether DMs reach the preferred level of confidence in their decisions (Hausmann & Läge, 2008). The appearance of an outlier can reduce a DM's confidence inability of acquired information to produce an accurate estimate. And this can trigger a search for additional expert forecasts. As a result the final aggregation weight of an outlier will be diminished by updating of the estimate update using new forecasts. A similar benefit from further evidence accumulation is not certain when DMs produce an estimate only at the end of a forecast sequence uture research should explore whether an outlier will encourage DMs to extend information search, and if the search will result in greater accuracy, both when the estimation is done step-by-step or at the end of a sequence. In the mean time, the results of this study recommend we simply write down all forecasts, and take a look at them simultaneously before making a decision.

3. Estimate revision with multiple advices

Abstract.

People, being egocentric, tend to make little use of an advice in order to revise their initial opinions. But with multiple advices, as results from the first experiment demonstrate, people will choose to revise more when they find their opinions to be outside a consensus. Analyses show that this consensus-dissensus category is a valid cue for an accuracy judgment. The second experiment examines whether concerns for rankings will make people make even more use of advices as previous research argues. The results do not support this hypothesis. Using data collected from two experiments, a simulation study suggests that having multiple advices, and its consequential revise-if-dissensus heuristic, can improve accuracy of revisions that decision makers may choose compared to having only a single advice.

3.1 Introduction

Often times decision makers (DMs) solicit advices from another person in order to confirm their initial opinions. Sometimes a received advice leads DMs to revise their initial position aggregating their own's and adviser's opinions. Combining estimates from multiple experts is shown to result in improved accuracy (Einhorn, 1972; Dawes & Corrigan, 1974; Doyle & Fenwick, 1976; Libby & Blashfield, 1978; Makridakis et al., 1982; Yaniv & Hogarth, 1993; Armstrong, 2001; Johnson et al., 2001; Budescu & Yu, 2007; Winkler & Clemen, 2004), and although giving all estimates an equal aggregation weight is shown to be the best combination scheme (Winkler, 1971; Newbold& Granger, 1974; Einhorn & Hogarth, 1975; Hogarth, 1978; Libby & Blashfield, 1978; Clemen & Winkler, 1986; Lawrence et al., 1986; de Menezes et al., 2000), this strategy is often ignored (Sniezek & Henry, 1989; Larrick & Soll, 2006). DMs are often judgmental, and how they combine their own estimates and advisors' depends on their subjective evaluation of those advices. This evaluation sometimes takes into account informative factors like advisors' reputation (Maines, 1996; Budescu et al., 2003), but sometimes it is biased by just the formats in which opinions are stated (Yaniv & Foster, 1995; also see Chapter 1).

Moreover, DMs are not estimator-agnostic. Research shows that DMs often give higher weights to their own estimates (for review see Bonaccio & Dalal, 2006) than to others. Even when assuming that DMs choose combination weights based on accuracy likelihood of estimates, over-weighting their own estimates is not unexpected, as DMs are generally found to be overconfident (Alpert & Raiffa, 1982; Brenner et al., 1996; Klayman et al., 1999; Soll & Klayman, 2004; Alicke & Govorun, 2005; Larrick et al, 2007; Soll, 2007) and confident DMs usually believe that their own opinions are superior to advisors' (Harvey & Fischer, 1997; Sniezek & Van Swol, 2001; Krueger, 2003; Gino & Moore, 2007; Soll & Larrick, 2009; Minson & Mueller, 2012). Such egocentric weighting has also been attributed to differential information regarding the access to the process that advisors use to produce estimates (Yaniv & Kleinberger, 2000; Yaniv, 2004a; Yaniv, 2004b), to the anchoring effect (Tversky & Kahnemann, 1974) where DMs do not technically combine estimates, but adjust their own initial estimates using information from advices (Lim & O'Connor, 1995; Harvey & Fischer, 1997). Producing estimates is not costless, it takes time, effort, and other resources. Such sunk cost (Arkes & Blummer, 1985) can make DMs reluctant to depart much from their own estimates, unless advices incur costs as well (Sniezek et al., 2004; Patt et al., 2006; Gino, 2008). When taking a closer look, DMs' egocentric revision is not a result of simply choosing higher weights to apply to their own opinions. Soll and Larrick (2009) found that when given another opinion, DMs choose between three revision choices of either to maintain their initial estimates, to switch completely to an estimate of an advisor, or to merely average the two estimates. The general egocentric overweighting is just a result of DMs' tendency to stick with their initial estimates.

Studies on revision self-weights mostly involve the case that DMs receive only a single advice, but it is not unusual that DMs consult two or more advisors. An economist might compare her GDP forecast with forecasts of multiple other economists. A mutual fund manager could read reports from a few analysts about the price of the stock she has just projected. Yaniv and Milyavsky (2007) found that DMs remained egocentric with their revisions as well after receiving multiple advices. In their studies, to produce revisions experimental participants seemed to discard an advice farthest from their initial estimates, and aggregate the rest. While Yaniv and Milyavsky (2007) consider that DMs revise by aggregating advices into their initial estimates with weights that are based on an egocentric distance, studies in this chapter will analyze DMs' revision rules when receiving multiple, specifically four, advices, based on the concept

proposed by Soll and Larrick (2009) but adapted to allow a treatment of multiple advices. Principally I argue that in the first stage of revision DMs first make a decision whether or not to revise, and this decision is subject to a pressure to conform to a group's opinion, and when DMs have decided to revise, in the second stage they will choose self-weights to combine their initial estimates with, as a less cognitively taxing strategy, a single advice representative of all advices.

The results from the first experiment supported the hypothesis that participants tended to revise more if their estimates did not conform to the consensus. That is when estimates appeared to be a dissensus lying outside the range that advices covered, participants were more likely to revise them; when estimates appeared to be a consensus, participants were more likely to retain their initial estimates. But the consensus-dissensus factor did not affect the level of self-weights they chose when they decided to revise. Furthermore, participants' revision decisions were largely not significantly affected by other factors examined here, which were an uncertainty level of the quantity to be estimated, and the information regarding their estimate ability relative to advisors. Analyses of accuracies revealed that dissensus estimates were inferior to advices; while consensus estimates were as accurate as advices so taking in advice would not result in significant accuracy improvement. This implied that participants' revision rule, particularly for the first stage of revision, was ecologically valid. This rule could be simplified as revise-if-dissensus (RiD) heuristic.

As researches in the field of finance have suggested that reputation concerns would increase the level of revision (Chavalier & Ellison, 1999; Hong et al., 2000; Clement & Tse, 2005), next I examined if a different reward scheme could nudge DMs to utilize advices more. Unlike the scheme of the first experiment whose rewards for revisions depended only on the accuracy of each independent revised estimate, the second experiment gave rewards based on accuracy rankings of revised estimates relative to other revisions. The results showed that participants remained sensitive to group conformity, and while the effect of rankings were significant for the first stage of revision decisions, participants did not retain their initial estimates less or apply higher self-weights to their revisions.

It is clear that for estimator-agnostic revision strategy like take-the-median, the more advices there are, the more accurate the revisions will be. However DMs show time and time again their resistance to using this strategy. Using estimates that participants from both experiments produced, I conducted a simulation study to

compare accuracies from different revision strategies. The results showed that, compared to the case of a single advice, when multiple advices were available so DMs could categorize their estimates as either a consensus or a dissensus and consequentially allowed the use of the RiD heuristic, one could expect DMs to arrive at a more accurate revision, especially if there were not overly egocentric in choosing self-weights.

3.2 Estimate revision

3.2.1 Revision process

3.2.1.1 Benefit of advice taking

Let Q be the quantity that a judge tries to estimate. This quantity cannot be predicted with a perfect accuracy, even when having all available information I and knowing the true information-processing model $\hat{M}(I)$, due to a random component ε with a zero mean and a variance of σ_ε^2, in other words $Q = \hat{M}(I) + \varepsilon$. Based on past experiences, a judge J has developed her own model to process information in order to produce E_J for the purpose of estimating Q, or $E_J = M_J(I)$. The judge is not perfect, her model does not always match that of the true model, missing it by m_J, that is $\hat{M}(I) = M_J(I) + m_J$. We can further assume that the judge is good enough at estimating Q, that is she is not biased, and over a large number of trials the judge's model on average matches the true model, that is m_J follows a zero-mean distribution. How good she is depends on the level of dispersion of her model's misses, indicated by s_J^2 ,the variance of m_J. So for each estimate, one can expect the (in)accuracy of the judge's estimate measured by a mean absolute deviation (*MAD*) from the correct value as the sum of the effects of the environment's randomness and of her model's miss, i.e. $MAD = \sqrt{(s_J^2 + \sigma_\varepsilon^2)\cdot 2\pi}$.[14] If a judge combines her own estimate with that of an advisor (whose model's misses has a variance of s_A^2) by using a self-weight w, her revised estimate will have an expected deviation of $\sqrt{(w^2 s_J^2 + (1-w)^2 s_A^2 + (2w^2 - 2w + 1)\sigma_\varepsilon^2)\cdot 2\pi}$. The *MAD* of the revision is smaller than that of her initial estimate as long as w is greater than $(s_A^2 - s_J^2)/(s_A^2 + s_J^2 + \sigma_\varepsilon^2)$ which is always smaller than 1. So there exists a self-

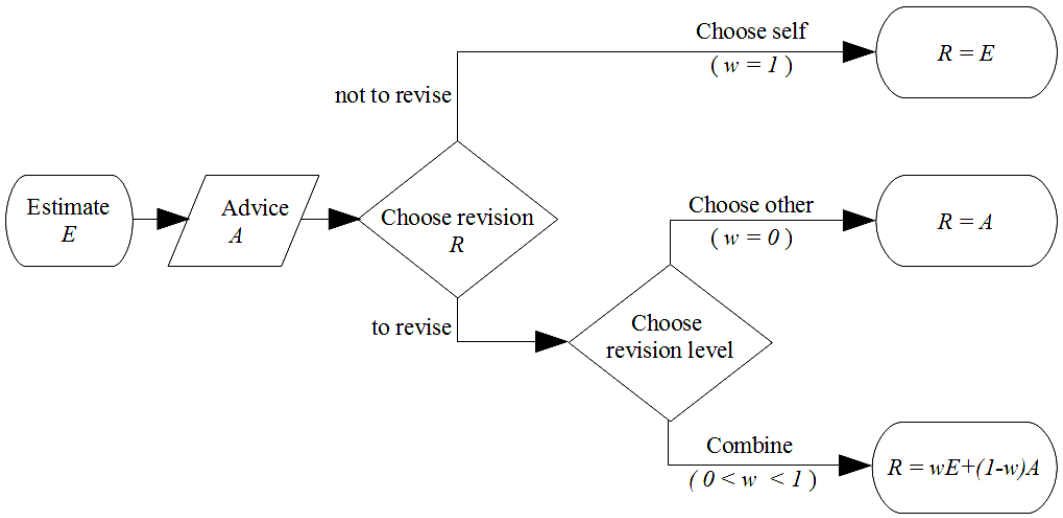
[14] Assuming both environment's errors and models' misses to be normally distributed.

weight below 1 with which a judge can use to combine her estimate and the advice together to produce a more accurate, i.e. with smaller *MAD*, estimate than her E_J .

3.2.1.2 Egocentric estimate revision

Previous research suggests that a judge gives a preferential consideration to her own estimate resulting in an under-adjustment with an average self-weight of approximately 0.7 (Lim & O'Conner, 1995; Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000; Yaniv, 2004a; Yaniv, 2004b). However this is based on the conventional approach in analyses of self-weights that assumes that DMs revise their estimates by choosing a level of self-weight along the 0-1 spectrum. Taking clues from the study of a two-cue prediction task by Lees and Triggs (1997) whose results showed a bimodal cue-weight distribution, Soll and Larrick (2009) argued that rather than choosing a self-weight, DMs instead would either choose to retain their initial estimates (w=1.0), to adopt a given advice (w=0.0), or to take an average of their estimates and an advice (w=0.5). The data from their experiments matched their argument, showing a W-shaped self-weight distribution with peaks at 0.0, 0.5, and 1.0. On average, participants were found to be egocentric not by choosing high self-weights but by choosing to retain their initial estimates at a high proportion, that is they were resistant to cognitive change (Greenwald, 1980). In fact this resulted in the average self-weight of 0.7, similar to other studies.

Figure 3.1. *Diagram of estimate revision processes*



Soll and Larrick (2009) did not necessarily argue that DMs consider the three revision choices at once. But if we are to apply further the egocentricity of DMs in the estimate revision process, it can be assumed that in the first step DMs choose to maintain their own estimates or not, that is choose between "self" (not to revise) or "not self" (to revise), and if they choose "not self" the next step is to choose a revision level, or a level of self-weight. This self-as-the-focus (Greenwald, 1980) nature means the process contains at least two steps rather than just one three-pronged step (Figure 3.1).

3.2.1.3 The case of multiple advices

When more than one advice is available, a judge could evaluate and decide how to use each advice in revision one-by-one, as assumed in the analyses by Yaniv and Milyavsky (2007). As DMs have generally been found to employ strategies that require less cognitive effort (Einhorn, 1971; Payne et al., 1988; Payne et al., 1993; Todd & Benbasat, 1994), it is likely that a judge will construct a *single* estimate that is representative of all advices, i.e. *other* estimates, and use that in revision processes instead. In fact research has demonstrated that DMs see themselves as being drawn from a different population from others (Harris & Guten, 1979; Weinstein, 1980; Perloff, 1982; Lehman & Nisbett, 1985; Perloff & Farbisz, 1985); moreover, as in stereotyping of an out-group (Boldry et al., 2007), a judge may also instinctively make a representative inference of the advisors' estimates. With that a judge considers only the *single-other* estimate(SO), the process in Figure 3.1 can be readily extended to deal with multiple advices. But from where is this SO approximated? Naturally, this will be the central value of the advice set, and the median is the likely inference for the central value (Yaniv, 1997; also see Chapter 1). So in this chapter, all analyses will consider the median advice as the SO of an advice set.

3.2.2 Decision factors in revision choice

3.2.2.1 Consensus-dissensus categorization

DMs seek advices to lower the uncertainty of, or to confirm, i.e. reduce some sort of internal conflict regarding, their initial thoughts (Festinger, 1954).However, with only a single advice and the egocentric tendency, rather than using that advice to re-evaluate their own initial thoughts, DMs seem to evaluate advice based on their own estimate instead. Researchers have found that a judge gives a higher weight to advices

that are consistent with their own estimates (Yaniv 2004a; Yaniv 2004b; Yaniv & Milyasky, 2009),which implied a distance effect, i.e. the greater distance between an advice and a judge's estimate is seen as an indicator of the advice's inaccuracy compared to her estimate but not of her estimate's inaccuracy compared to the advice.

But the situation is likely to be different if a judge receives more than one advice. Multiple advisors could create a greater pressure for a judge to reconsider her initial estimate, as DMs also have a penchant to conform to a group's majority opinion (Asch, 1951). Multiple advices give a context that a judge could directly perceive how much her estimate conforms to the group; for example, with a few advices, a judge's estimate can be clearly categorized as either a consensus, i.e. that lies outside inside an advice set, or a dissensus, i.e. an estimate that lies outside an advice set.[15] When her estimate appears to be a dissensus, the deviation from an advice set accentuates the conflict, and consequently urges the judge to revise; while this urge is likely to be absent when her estimate is a consensus. Also due to preference for consonant information (Frey, 1981), that is DMs utilize advices more if they are agreeing to one another (Yaniv et al., 2009), when her estimate is a dissensus, an advice set could be seen as more cohesive, and when her estimate is a consensus the advice set could be seen as more disagreeing. As a dissensus estimate is farther from the SO than a consensus estimate, given the same advice set, this hypothesis in a sense implies that a judge will incorporate advices more if they lie further away from her own estimate, opposite of the distance effect.

3.2.2.2 Uncertainty of the estimation environment

As demonstrated by Larrick and Soll (2009)'s study, under-revision of an estimate is a result of a judge's disproportionately high tendency to choose self. Generally DMs in many situations are reluctant to take even an advantageous action, due to the status quo bias (Samuelson & Zeckhauser, 1988), the omission bias (Ritov & Baron, 1992), or conservatism (Harvey & Harries, 2004). Although inaction can be an optimal strategy if the expected benefit of an action does not exceed its cost (Beach & Mitchell, 1978; Christensen-Szalanski, 1978; Payne, et al. 1993). Similarly a judge will not be compelled to revise if she does not expect that incorporating an advice into a

[15] In this chapter an estimate is considered to lie outside an advice set if it is either lower than the lowest advice or higher than the highest advice. For example, let an advice set be {5, 8, 10}, if the judge's estimate is 3, then it is a dissensus. The estimate of 12 is also a dissensus. If her estimate is 7, it is a consensus. So is it if her estimate is 10 or 5.

revision will result in a sufficient benefit that is accuracy improvement, compared to costs associated with revision, such as a cognitive cost from re-calculating an estimate, or a psychological cost of diverting from egocentric confidence. If a judge expects that the revised estimate has a variance of s_R^2 (where $s_R^2 < s_J^2$, or else a judge would not revise), the revision's *MAD* is $\sqrt{(s_R^2+\sigma_\varepsilon^2)\cdot 2\pi}$; while the no-revision's *MAD* is $\sqrt{(s_J^2+\sigma_\varepsilon^2)\cdot 2\pi}$. The relative benefit of the revision depends on the ratio of two *MAD*s or $\sqrt{(s_R^2+\sigma_\varepsilon^2)/(s_J^2+\sigma_\varepsilon^2)}$. Since $s_R^2 < s_J^2$, this relative benefit decreases as the environmental uncertainty increases. So one could expect that a judge is less likely to revise her initial thought if she thinks it is not likely that she will make an accurate estimation.

3.2.2.3 Confidence of a judge

As the aim of advice taking is to improve accuracy by reducing the expected deviation from the true value, normatively a judge should give combination weights to each estimate proportionate to expected accuracies of their estimators. A judge who believes that she is better at the estimate task than advisors will then apply a higher self-weight to her own estimate when taking and combining advices than a judge who believes otherwise. In fact, in the experiment by Yaniv (2004), participants that assumed to be more knowledgeable about the topic questioned chose higher self-weights than less knowledgeable participants.

3.3 Experiment 1

In order to test the effects of the above hypothesized factors, I conducted the following experiment, in which participants were tasked to produce estimates of a house price in two simulated markets, one with a high and the other with a low uncertainty (or price-variance) level. For each estimation participants would be given a set of three advices, from which participants could clearly separate their initial estimates into either a consensus or a dissensus with regard to the advice set. Participants would also be primed with information of their performance compared to other participants whose initial estimates would be in the pool that advices would be drawn from.

3.3.1 Method

Procedures. The experiment was conducted in a laboratory in Barcelona in Spanish on personal computers. Participants were told that they would participate in two consecutive rounds of similar experiments. Their task was to predict a sale price of a series of houses, in a unit of one thousand U.S. dollars, based on the following four characteristics: 1) the number of bedrooms, 2) the distance between a house and the city center, 3) whether a house has a swimming pool, and 4) the year that a house was constructed or received a major renovation. They were told that the data were from the actual houses that were sold during the year 2000, and that all houses in the series in the same round were from the same city (i.e. the same housing market). Each round contained four phases. 1) In the first phase, participants learned by observing the relationship between prices and the four characteristics of 15 houses, shown in 3 blocks with 5 houses per block, each block was on screen for 25 seconds. 2) In the second phase, participants practiced by producing price estimates of 7 houses, one at a time, based on the given four characteristics of a house. After the four characteristics appeared on screen they had 20 seconds to give an estimate, after which the correct price would appear as a feedback. 3) In the third phase, the procedure was similar to the second phase but without the correct-price feedbacks. At the end of the phase they were shown how they had performed in a form of a percentile compared to other participants in the laboratory. A brief description of how to interpret their percentiles was also on screen, (they were also told how the percentile would be calculated, and of the meaning of the percentile data in the instructions). The participants were told that rewards would be given in this phase based on the accuracy of each individual estimate: when the difference between an estimate and the real price was 2 units or fewer, they would be rewarded 10 Euro cents; when the difference was higher than 2 but lower or equal to 7, the reward was 5 Euro cents; when the difference was higher than 7 but lower or equal to 15, the reward was Euro 2 cents; and when the difference was more than 15, they would get no rewards. 4) The fourth phase was the estimation-then-revision phase with a series of 10 houses. In the first half of each estimation, participants were shown the four characteristics of a house and asked to produce the estimate of its sale price within 20 seconds (a counting-down clock was on the top right corner of the screen, but the time limit was not enforced). In the second half, each participant received three advices randomly selected from the estimates to the same house produced by other participants, as stated in the instructions. The advices were shown on the screen in a vertical row under the participant's own estimate. Participants were asked to give a revision within

20 seconds (the time limit was not enforced). Participants were told that rewards would be given for each individual initial estimate and revision. The instruction gave details of the reward scheme which was the same as the third phase's, but with the reward amounts five times higher in all steps. Upon finishing the first round before the second round started, participants were told that that data for the following round came from the same year as in the first round but from a different city (different housing market).

Stimuli. There were two parts in each item: 1) the base data, contained the house characteristics and the base house prices, 2) the random component, to be added to the base price to produce the real price. The set of four characteristics were simulated individually and independently. That a house had a pool or not was simulated using a binomial distribution, with the likelihood of having a pool at 0.7. The other three characteristics were simulated following a uniform distribution rounded to the nearest integers. The numbers of bedrooms were between 3 and 6, the distance to the city center between 1 to 30 kilometers, the year of the house's construction or major renovation between 1970 and 1990. The basic house prices were calculated by the following "true model" which yielded the variance of 1618.2,

$$House\ \ price = 200 + (6 \cdot Bedroom) + (-1 \cdot Distance) + (12 \cdot Pool) + (Year - 1970)$$

The random error components for the high-uncertainty condition were selected from a normal distribution with a variance of 1078.8, so that the true model captured 60% of the price variation. Those for the low-uncertainty condition were set to let the true model capture 85% of the price variation by being selected from a normal distribution with a variance of 286.6.

For the base data, six data sets were generated: set 1 and 2 contained the data of 29 houses, and set 3 to 6 contained the data of 10 houses each. Two sets of high-uncertainty error components were generated: set 1 with 29 data points, set 2 with 10 data points; the same amount of data points were also generated from low-uncertainty error components. The matching of the stimuli sets and participant groups in each round and each phase was shown in Table 3.1.

Table 3.1. *Matching of stimuli sets and participant groups*

		Participant group 1	**Participant group 2**	**Participant group 3**	**Participant group 4**
Round 1	**Phase 1-3**	Base data set 1 High-error set 1	Base data set 1 High-error set 1	Base data set 1 Low-error set 1	Base data set 1 Low-error set 1
	Phase 4	Base data set 3 High-error set 2	Base data set 4 High-error set 2	Base data set 5 Low-error set 2	Base data set 6 Low-error set 2
Round 2	**Phase 1-3**	Base data set 2 Low-error set 1	Base data set 2 Low-error set 1	Base data set 2 High-error set 1	Base data set 2 High-error set 1
	Phase 4	Base data set 5 Low-error set 2	Base data set 6 Low-error set 2	Base data set 3 High-error set 2	Base data set 4 High-error set 2

Participants. Sixteen participants aged between 19 and 23 years took part in this experiment, recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the experimental laboratory. Ten were female, six were male. They were randomly separated into four groups of equal size for the matching with stimuli as described in Table 3.1. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 10.20 Euros.

3.3.2 Results

Prior to the analyses, some of the response items were discarded due to various reasons. First were items that contained extreme outlying initial estimates, there were probably typing errors caused by time-limit pressure.[16] Next are those whose revisions were based on an advice set that had at least one extreme outlying advice since they could affect revision choices differently than the rest.[17] As participants were permitted to answer only in integer, I included only response items whose gaps between the initial estimates and the advices' medians were 2 or larger, to allow only revisions whose choices included a combination between advices and participants' own estimate.[18] In addition, I excluded response items whose revisions were outside the ranges that initial estimates of both judges and advisors together covered.[19] Such revisions could be

[16] These were 4 items all from the high-variance condition, and none from the low-variance condition.

[17] There were 9 items, all from the high-variance condition.

[18] 8 and 22 items in the high- and low-variance conditions respectively were discarded.

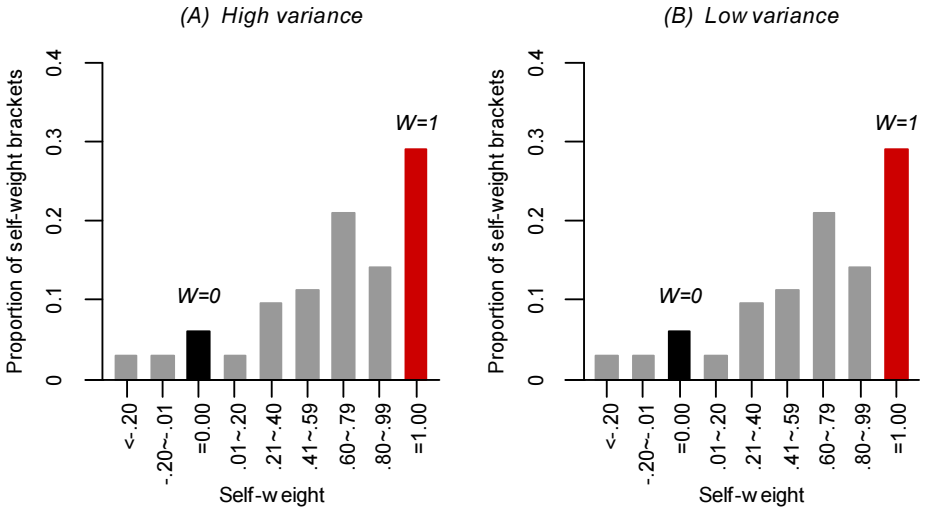
[19] There were 7 and 5 items in the high- and low-variance conditions respectively.

thought of as not taking advices, or in some case even participants' own initial estimates, into account. Among data items that remained for the analyses, under the high-variance condition there were 64 whose initial responses from the judges were consensus estimates, and 70 whose initial responses were dissensus estimates; under the low-variance condition there were 60 items with a consensus initial estimate, and 73with a dissensus initial estimate. Participants were also labeled as one of the two categories based on the information of comparative performance during the test phase of each round. Those with a percentile over 50 were top performers for that round, and those with a percentile below 50 were bottom performers for that round.

3.3.2.1 Revision patterns

Self-weight distributions. Distributions of revision self-weights, computed by $w = (R - C)/(E - C)$,[20] displayed multiple peaks resembling the results found by Soll and Larrick (2009) (Figure 3.2).

Figure 3.2. *Distribution of self-weights by variance levels, Experiment*

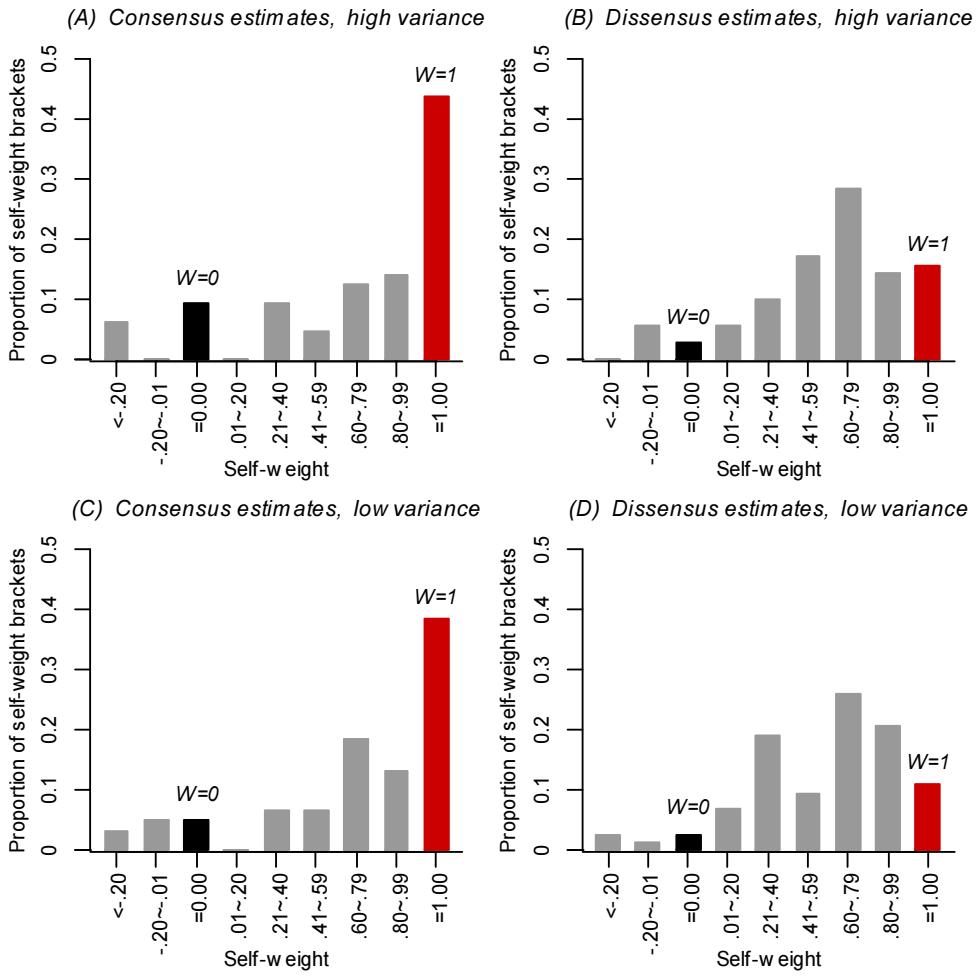


These distributions suggested that participants might have considered at least initially between maintaining their initial estimates (self-choosers) or incorporating advisors' opinions (revisers). When examining the distributions of self-weight by types of initial

[20] Strictly, this is not a real self-weight since the true SOs are unknown. Since the SOs are likely to center around the advices' medians, this type of self-weights is the "expected" self-weights.

estimates (Figure 3.3), it is clear that the participants felt more compelled to revise when their initial estimates were dissensus. The following analyses will concern two revisions choices: to choose self or not, and if choose to revise, what level of self-weight to use.

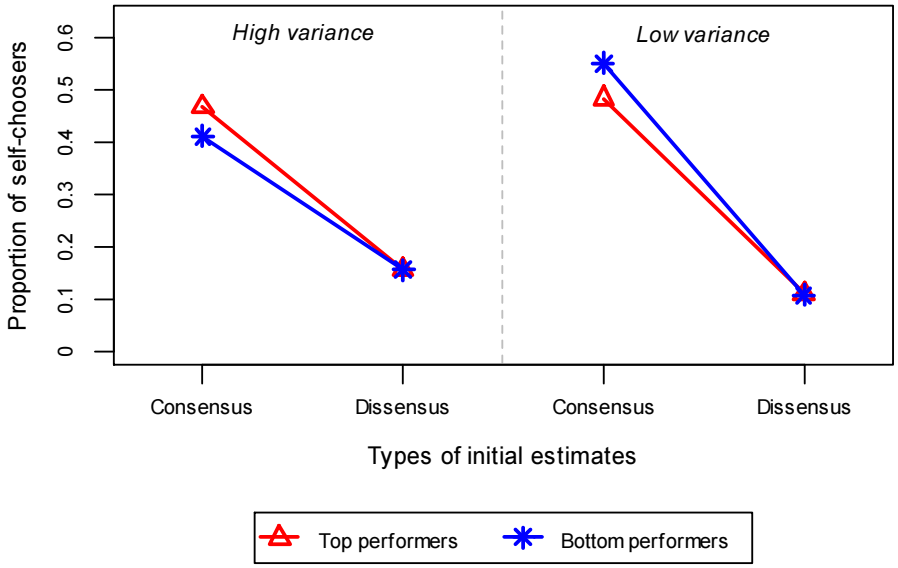
Figure 3.3. *Distribution of self-weights by variance levels and types of initial estimates, Experiment 1*



Choosing self. The generalized estimating equations (GEE) were used in the analyses due to the fact that each one participant contributed multiple responses. The results of logit regression revealed that only the main effect of the consensus-dissensus categorization of initial estimates was significant ($p<.01$), but not the effects of the market uncertainty ($p\approx.32$) or of comparative performance information ($p\approx.67$); none of the interaction effects among the three factors was significant either.

Figure 3.4 shows, as in Figure 3, that participants were significantly more likely to choose-self when their initial estimates were consensus than when they were dissensus (M=0.48 vs. 0.13, p<.01), in fact this is true in both the high- (M=0.44 vs. 0.16, p<.001) and the low-variance conditions (M=0.52 vs. 0.11; p<.001). While they were equally likely to choose-self when the price variance was high than when it was low (M=0.29 vs. 0.29, p≈.32), participants were less likely to choose-self in the high- than in the low-variance market if their initial estimates were of consensus (M=0.44 vs. 0.52, p≈.36); but among dissensus estimates, participants were more likely to choose-self in the high- than in the low-variance market (M=0.16 vs. 0.11, p≈.29).

Figure 3.4. *Proportions of choose-self revisions, Experiment 1*



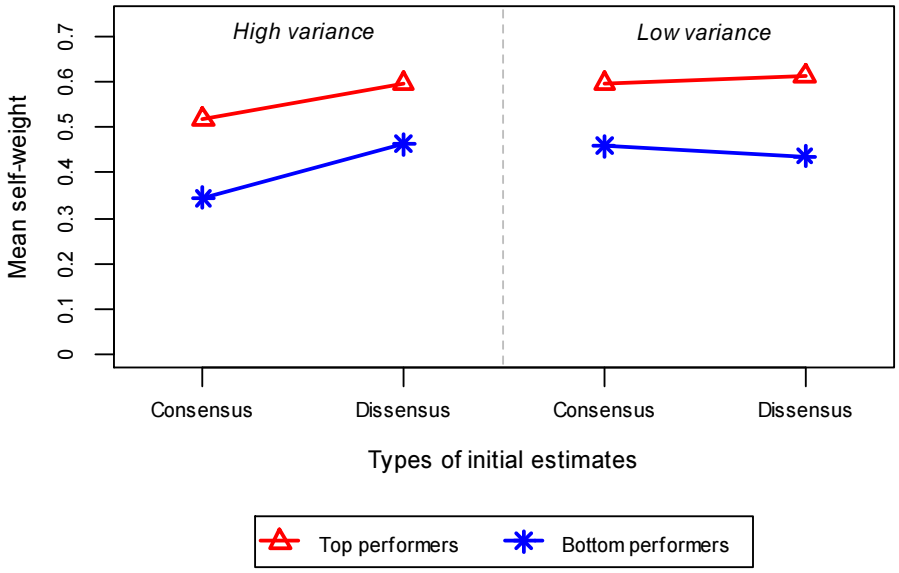
As expected, top performers were more, but not significantly, likely to choose-self than bottom performers (M=0.59 vs. 0.42, p≈.67). But in all sub-conditions, this pattern did not always show, and the effect was not significant (M=0.47 vs. 0.41, p≈.66 in the high-variance consensus-estimate; M=0.16 vs. 0.16, p≈.37 in the high-variance dissensus-estimate; M=0.48 vs. 0.52, p≈.61 in the low-variance consensus-estimate; M=0.11 vs. 0.11, p≈.95 in the low-variance dissensus-estimate conditions).

Revisers' self-weights. The results from the GEE regression of self-weights showed that past performance information was the only significant main effect (p<.05),

but not a consensus-dissensus nature of initial estimates (p≈.88) or the market's uncertainty levels (p≈.35); none of the interaction effects among the three factors was significant either.

As shown in Figure 3.5, overall consensus estimates received lower revising self-weights than dissensus estimates did, but not significantly (M=0.47 vs. 0.53, p≈.13). The same pattern was found in the high-uncertainty condition (M=0.42 vs. 0.53, p≈.18), but in the low-uncertainty condition the pattern was the opposite (M=0.53 vs. 0.52, p≈.32). Participants applied lower self-weights when the uncertainty of the market was high than when it was low (M=0.49 vs. 0.53, p≈.48). The pattern was found when analyzing only consensus estimates (M=0.49 vs. 0.53, p≈.16), however with dissensus estimates, the high-variance condition saw discernibly higher self-weights than the low-variance condition did (M=0.53 vs. 0.52, p≈.87).

Figure 3.5. *Mean self-weight in revisions among revisers' responses, Experiment 1*



Overall top performers revised their initial estimates with significantly higher self-weights than bottom performers (M=0.59 vs. 0.43, p<.01).While the pattern was consistent in all four conditions, none was significant (M=0.52 vs. 0.35, p≈.15 in the high-variance consensus-estimate; M=0.60 vs. 0.46, p≈.11 in the high-variance dissensus-estimate; M=0.60 vs. 0.46, p≈.89 in the low-variance consensus-estimate; M=0.61 vs. 0.44, p<.10 in the low-variance dissensus-estimate conditions).

3.3.2.2 Revision cue validity and revision accuracy

Revision cues. The results above that participants decided whether to revise their initial estimates at all depended on whether those estimates were consensus or dissensus, implying that participants perceived such category as a signal of accuracy, particularly in comparison to the SO of the advice sets. This accuracy cue was indeed valid as we can see in Table 3.2 that lists *inaccuracy* of estimates measured by the mean absolute deviations (MADs) between estimates and the true-model values $\hat{M}$. The reason that I did not use "correct" sales prices because the judge is supposed to predict the true model's quantity rather than to attempt to capture also the random component of a market included in the sales prices.

Table 3.2. *Mean absolute deviations of initial estimates, and advices*

	High variance		**Low variance**	
	Initial estimate	**SO**	**Initial estimate**	**SO**
Consensus				
Top	6.17	8.23	6.45	7.55
Bottom	9.41	7.88	8.83	7.93
All	7.89	8.05	7.60	7.73
Dissensus				
Top	14.34	8.97	9.42	8.64
Bottom	19.00	8.00	16.08	6.05
All	16.47	8.53	12.79	7.33

MADs of consensus estimates were significantly lower than MADs of disensus estimates in both the high- (MAD=7.89 vs. 16.47, p<.001, Mann-Whitney test), and low-variance conditions (M=7.60 vs. 12.79, p<.01, Mann-Whitney test). And more importantly, dissensus estimates deviated significantly further from the true-model prices than the SOs did in both the high- (MAD=16.47 vs.8.53, p<.01, Wilcoxon test), and low-variance conditions (MAD=12.79 vs.7.33, p<.001, Wilcoxon test); while the accuracy advantage of consensus estimates over their respective SOs was only slight

(MAD=7.89 vs.8.05, p≈.95 in the high-variance condition, Wilcoxon test; MAD=7.60 vs.7.73, p≈.96 in the low-variance condition, Wilcoxon test).

According to the results of the experiment, participants also took into consideration their performance during the test phase that top performers decided on higher self-weights when revising estimates than bottom performers. This implied that top performers were more confident in their own estimates over advices than bottom performers were. I compared accuracies of estimates versus accuracies of the SOs using the mean differences in absolute deviations from the true-model prices of estimates and of their respective SOs (or MDADs, which in this case is basically MADs of the estimates minus MADs of the SOs). Among consensus estimates, top performers was superior with lower MDAD than bottom performers, but not significantly (MDAD=-2.07 vs. 1.53, p≈.25 in the high-variance condition, Mann-Whitney test; MDAD=-1.10 vs. 0.90,p≈.57in the low-variance condition, Mann-Whitney test).Among dissensus estimates, the accuracy advantage of top performers was also greater than bottom performers but not always significantly (MDAD=5.37 vs. 11.00, p<.10 in the high-variance condition, Mann-Whitney test; MDAD=7.78 vs. 10.03,p<.001 in the low-variance condition, Mann-Whitney test).

Revision accuracies. While being a consensus signaled a superior accuracy, and almost half of the times experimental participants' decided to retain their consensus estimates, it did not mean that no significant improvement could be had from revisions. To explore this, I compared accuracies in terms of MADs of estimates and revisions with a non-egocentric take-the-median (TTM) that has been shown to be an often used aggregation strategy that can produce a fairly accurate estimate (Yaniv, 1997; Harries et al., 2004; also see Chapter 1).

As seen in Table 3.3, among consensus initial estimates, the accuracies of those that participants decided not to revise were better than those that participants revise, but not to a significant degree (MAD=6.61 vs. 8.89, p≈.11 in the high-variance condition, Mann-Whitney test; MAD=7.39 vs. 7.83, p≈.92 in the low-variance condition, Mann-Whitney test). Moreover, both self-choosers and revisers were only slightly less accurate when compared to TTM in both high- (MAD=6.61 vs. 6.43, p≈.84 for self-choosers, Mann-Whitney test; MAD=8.89 vs. 7.33, p≈.27 for revisers, Mann-Whitney test), and low-variance conditions (MAD=7.39 vs. 6.55, p≈.57 for self-choosers, Mann-Whitney test; MAD=7.83 vs. 5.69, p≈.13 for revisers, Mann-Whitney test). As with not much room for improvement, revisers' final estimates showed only small gains in

accuracies, not significantly different from initial estimates (MAD=8.53 vs. 8.88, p≈.56 in the high-variance conditions, Mann-Whitney test; MAD=5.72 vs. 7.83, p≈.12 in the low-variance conditions, Mann-Whitney test).

Table 3.3. *Mean absolute deviations of estimates, advices, and revision strategies.*

	High variance			Low variance		
	Initial estimate	**Revision**	**TTM**	**Initial estimate**	**Revision**	**TTM**
Consensus						
Self-chooser	6.61	n/a	6.43	7.39	n/a	6.55
Reviser	8.89	8.53	7.33	7.83	5.72	5.69
Dissensus						
Self-chooser	14.36	n/a	5.95	6.38	n/a	4.88
Reviser	16.86	8.64	7.82	13.58	7.95	6.29

Among dissensus estimates, the levels of accuracy did not differ between self-choosers and revisers in the high-variance condition (MAD=14.36 vs.16.86, p≈.67, Mann-Whitney test),but in the low-variance condition self-choosers exhibited significantly better accuracy (MAD=6.38 vs.13.58, p<.05, Mann-Whitney test). Comparing to TTM, participants' initial estimates were significantly less accurate in both high- (MAD=14.36 vs. 5.95, p<.05 for self-choosers, Mann-Whitney test; MAD=16.86 vs. 7.82, p<.01 for revisers, Mann-Whitney test), and low-variance conditions (MAD=16.86 vs. 7.82, p<.01 for self-choosers, Mann-Whitney test; MAD=13.58 vs. 6.29, p<.001 for revisers, Mann-Whitney test). Those dissensus estimates that were revised, accuracies of revisions improved significantly from their initial figures (MAD=8.64 vs. 16.86, p<.01 in the high-variance condition, Mann-Whitney test; MAD=7.95 vs. 13.58, p<.001 in the low-variance condition, Mann-Whitney test.). In fact, while accuracies of revisions were worse than TTM, the differences were not significant (MAD=8.64 vs. 7.82, p≈.68 in the high-variance condition, Mann-Whitney test; MAD=7.95 vs. 6.29, p<.10 in the low-variance condition, Mann-Whitney test).

3.3.2.3 Discussion

The results from this experiment clearly suggested that having multiple advices could pressure DMs into incorporating others' opinions more, specifically, when DMs' own initial estimates appear as a dissensus.[21] Indeed from this experiment, being a dissensus appeared to be a valid reason to revise, as such estimates proved to be significantly inaccurate when compared to estimates from advisors. On the other hand multiple advices could also adversely give DMs the reason to favor a non-revision if an initial estimate is a concensus.[22] But analyses also showed that such estimates were already as accurate as advices, in fact the accuracies were not significantly different than if participants had decided to take the non-egocentric medians.

While a consensus-dissensus category mattered significantly at the first stage, it did not at the second stage where self-weights were chosen for revisers. In fact, among dissensus initial estimates that were adjusted only 52.5% and 40.0% of in the high- and low-variance conditions respectively were revised enough to become consensus to given advices. That is not being a consensus was a reason to revise but being a consensus was not a revision goal. For revisers, participants might have just combined estimates as in the model proposed by Soll and Larrick (2009).

Additional two factors, information about comparative estimate ability, and levels of price uncertainty, were also examined. From the results, participants who were aware of their higher expected accuracy appeared to use higher self-weights than others. But I did not find a low level of revision when an uncertainty level was high as I hypothesized. However it was possible that the difference in uncertainty levels was too narrow to generate a significant impact.

Overall, the strategy used by participants in this experiment could be summarized as a revise-if-dissensus (RiD) heuristic, where DMs adhere to their initial estimates if they are a consensus, and combine their initial estimates and the SOs using some level of self-weight when those initial estimates are a dissensus.

3.4 Reputation competition and herding

[21] Revisions in this case featured self-weights (self-choosers and revisers together) significantly lower than 0.7 ($M=0.61$, $p<.05$ in the high-variance condition, Mann-Whitney one-sided test for a mean of 0.7; $M=0.58$, $p<.001$ in the low-variance condition, Mann-Whitney one-sided test for a mean of 0.7).

[22] Revisions for consensus estimates featured self-weights (self-choosers and revisers together) as high as 0.7 ($M=0.67$, $p\approx.19$ in the high-variance condition, Mann-Whitney one-sided test for a mean of 0.7; $M=0.77$, $p<.01$ in the low-variance condition, Mann-Whitney one-sided test for a mean of 0.7).

The normative analysis by Froot et al. (1992) showed that in speculative trading when the investment time horizon was short, herding (i.e. adopting a consensus opinion) was the equilibrium, as traders tried to infer information from trades that other had made. This, the authors argued, would lead to inefficiencies as the market as a whole used too much of some types of information and too little, or even none, of others. The decision to herd might be different between DMs of different competence. In the model by Trueman (1994), two analysts made and announced earnings forecasts based on private information. However, when the more competent analyst made the announcement first, in which the forecast was optimistic, the less competent analyst would be more likely to adjust the forecast in line with that first announcement.

DMs' concern for their reputations has also been suggested as a reason for herding, and indeed financial analysts surveyed by Brown et al. (2013) claimed that their standings in analyst rankings were much more importation than the accuracy of their forecasts. Scharfstein and Stein (2001) developed a model with two managers making decisions on a corporate investment using a market signal and each other's decision. When the sole objective was a return on investment, one manager's investment decision would be based on information inferred from the other manager's decision as well as the information from the market signal. But when the concern about the reputation ("smart" versus "dumb") was in the equation, the equilibrium dictated that each manager always mimicked investment decision of the other manager. Studies of the field data concerning decisions made by financial professionals showed herding as a result of a career concern, i.e. a practical form of a reputation concern. Chevalier and Ellison (1999) analyzed the portfolio choices of mutual fund managers during the years 1992-1994 and found that younger managers were more likely to avoid unsystematic risk by constructing their portfolios in a more conventional manner, i.e. following the mean industry sector weighting beta. They suggested that such herding among younger managers were the result of a career concern, since after controlling for the fund performances, young managers whose choices deviated more from the unsystematic risk level in their objective group were more likely to lose their employment. Hong et al. (2000) also found that analysts faced a grimmer probability in their career prospect if they had had a poor forecast performance in the past especially when that forecast deviated widely from the consensus. As a result, inexperienced analysts were less likely to be the first to issue

a forecast, more likely to revise their forecast multiple times, and their forecasts tended not to deviate from the consensus. However when Clement and Tse (2005) examined financial analysts' annual earnings forecasts during the years 1989 to 1998, they found the motivation for a bold forecast revision in factors other than a career concern. Factors associated with boldness, e.g. a size of an analyst's brokerage firm, did not predict the job termination after making a bold forecast, while factors that predicted the job termination had an opposite effect, e.g. analysts who followed a large number of firms or industries were often safe after making a bold forecast, but these analysts were also more likely to make a forecast close to a consensus.

3.5 Experiment 2

This experiment investigates if reputation competition will encourage DMs to revise more, i.e. to choose-self less or to apply lower self-weights to revisers, compared to results from Experiment 1. While DMs' reputation regarding estimate ability is complex involving both independent, i.e. accuracy of a judge herself, and comparative, i.e. accuracy ranking among other estimators. In this experiment I will focus only on the effect of a concern on ranking on revision decisions.

3.5.1 Method

Procedures and stimuli. This experiment followed the procedures exactly as those of Experiment 1, with the same data sets for house prices and house characteristics. The only difference was that in this experiment participants were told that the rewards for the revisions in the estimation phase would depend on the rankings of how accurate their revisions were compared to the group of those four participants with whom they exchanged the estimates. For revisions, the reward of 40 Euro cents would be given to a participant who produced the most accurate estimates among the group; the second most accurate participant would receive 20 cents; the third most accurate participants would receive 10 cents; the least accurate participants would receive no reward.

Participants. Sixteen participants aged between 18 and 28 years took part in this experiment, recruited via emails from the pool of undergraduate students at Universitat Pompeu Fabra registered with the experimental laboratory. Eight were female, eight

were male. Participants received a participation fee of 3 Euros plus a performance-based reward. The mean remuneration was 12.28 Euros.

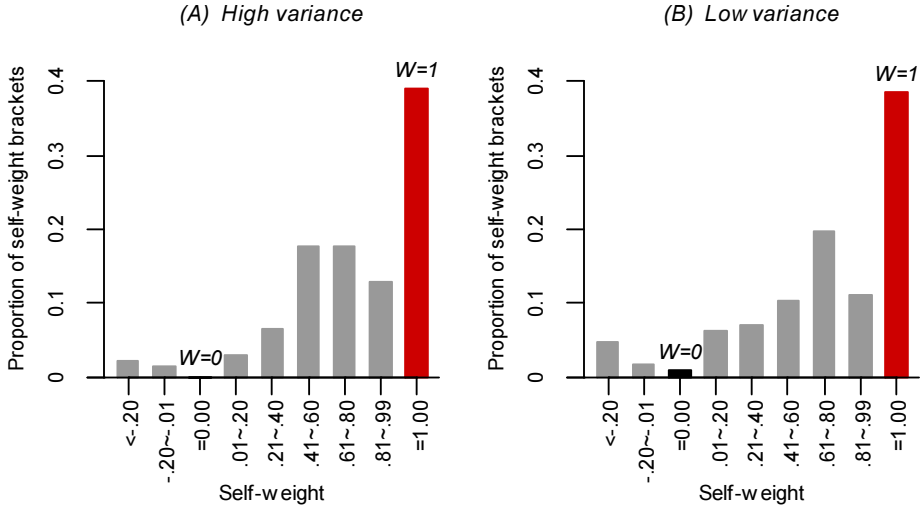
3.5.2 Results

Some responses items were discarded according to the same criteria as in the first experiment,[23] after which in the high-variance condition there were 68 consensus initial estimates, and 77 dissensus initial estimates left; in the low-variance condition there were 62 consensus initial estimates, and 71 dissensus initial estimates left.

3.5.2.1 Revision patterns

Self-weight distributions. The revision patterns were largely similar to those in the previous experiment. Revision choices depicted in Figure 3.6 also suggested that in this experiment revision decisions similarly appeared to be a two-step process of choosing whether to revise or not, and then if choosing to do so, how much to revise.

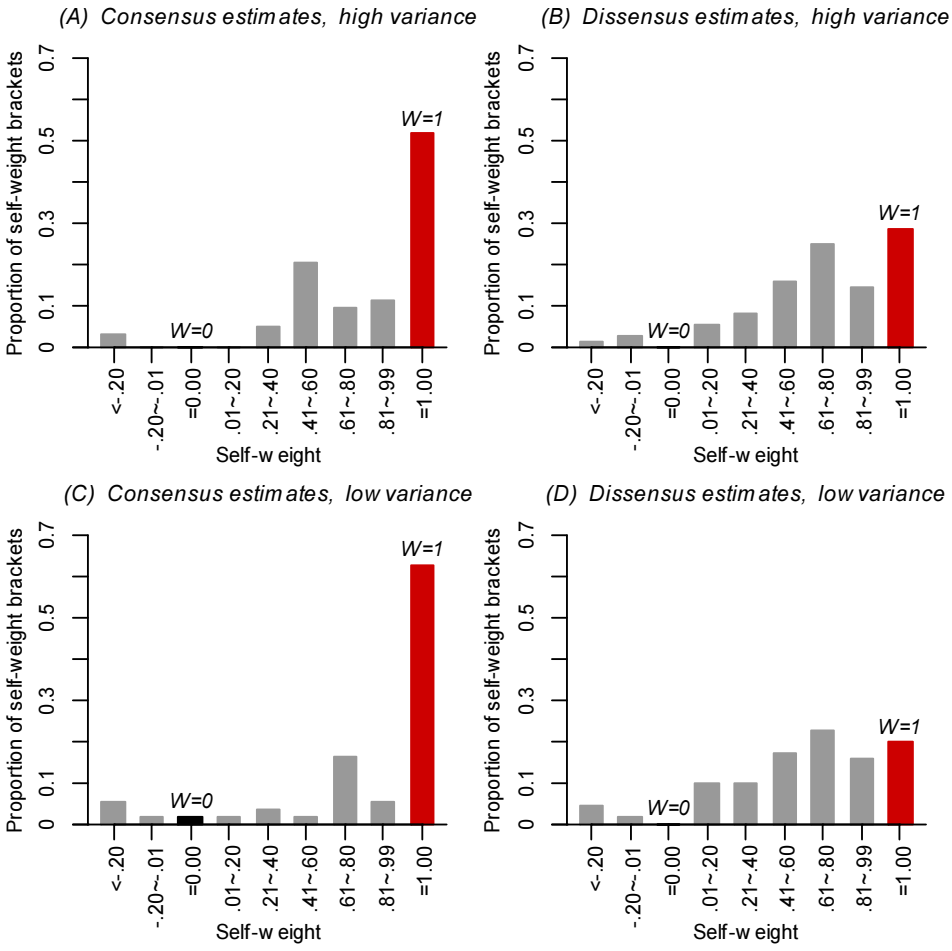
Figure 3.6. *Distribution of self-weights by variance levels, Experiment 2*



[23] There was no apparent outlier. I discarded 13 and 19 items from the high- and low-variance conditions that had initial estimates with a distance of less than 2 from the SO. Additional 2 and 8 items were discarded from the high- and low-variance conditions respectively due to that their revisions lay outside the range covered by all of the item's estimates. Lastly, 4 and 6 items were discarded from the high- and low-variance conditions respectively as their self-weights were higher than 1.

Figure 3.7 demonstrated that that an estimate was a consensus or a dissensus mattered in whether participants would choose to maintain their initial judgment.

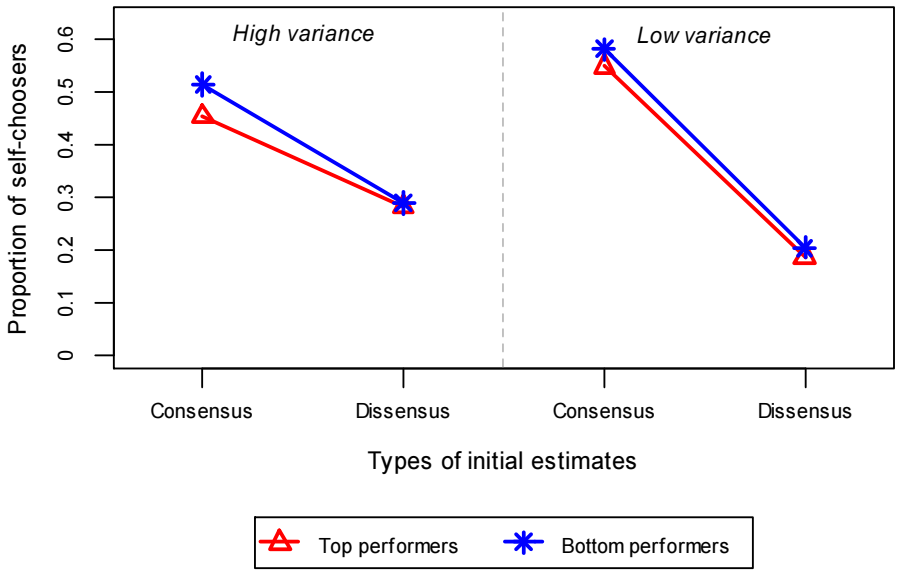
Figure 3.7. *Distribution of self-weights by variance levels and types of initial estimates, Experiment 2*



Choosing self. The results of GEE logit regression revealed that only the main effect of initial estimates' consensus-dissensus category was significant ($p<.001$), but not the effects of the market uncertainty ($p\approx.62$) or of information of past comparative performance ($p\approx.73$); none of the interaction effects among the three factors was significant either. Figure 3.8 shows, as in Figure 3.7, that participants were significantly more likely to choose-self when their initial estimates were consensus than when they were dissensus (M=0.52 vs. 0.24, $p<.001$), in fact this is true in both the high- (M=0.49

vs. 0.29, p<.01) and the low-variance conditions (M=0.57 vs. 0.20; p<.001). The difference of likelihoods to choose-self when the price variance was high and when it was low was negligible (M=0.38 vs. 0.37, p≈.91). When their initial estimates were consensus participants were less likely to choose-self in the high- than in the low-variance (M=0.49 vs. 0.57, p≈.54), but among dissensus estimates, the result was the opposite (M=0.29 vs. 0.20, p≈.27). The effects were not significant in either case.

Figure 3.8. *Proportions of choose-self revisions, Experiment*

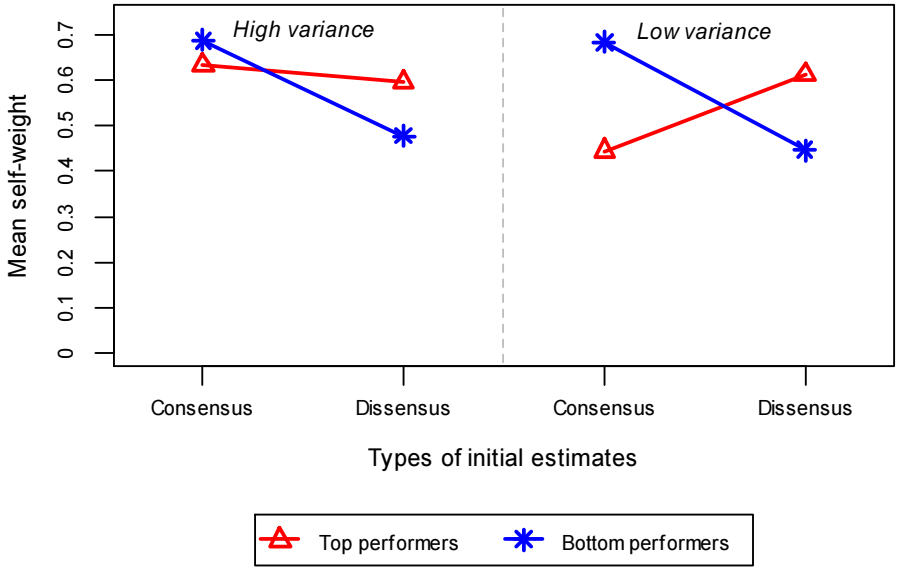


Opposite of what expected, top performers were slightly less likely to choose-self than bottom performers (M=0.36 vs. 0.38, p≈.71), and the pattern remained in all sub-conditions (M=0.46 vs. 0.51, p≈.67 in the high-variance consensus-estimate; M=0.28 vs. 0.29, p≈.99 in the high-variance dissensus-estimate; M=0.55 vs. 0.58, p≈.64 in the low-variance consensus-estimate; M=0.19 vs. 0.21, p≈.83 in the low-variance dissensus-estimate conditions).

Revisers' self-weights. The results from the GEE regression showed no significant main effect of any of the three hypothesized factors (p≈.17 for consensus-dissensus effect; p≈.98 for the market uncertainty effect; p≈.32 for the information of the past performance). The only significant interaction effect came from that between the consensus-dissensus category and the past-performance information effects. Overall, consensus estimates received higher, self-weights, but not significantly so, when revised

than dissensus estimates did (M=0.62 vs. 0.54, p≈.29).This was also the case for both the high-(M=0.66 vs. 0.56, p≈.37), and the low-uncertainty conditions (M=0.56 vs. 0.51, p≈.68). Participants used higher, albeit not significantly, self-weights in the high- than in the low-variance conditions (M=0.60 vs. 0.53, p≈.98). The same pattern was present both when considering only consensus estimates, (M=0.66 vs. 0.56, p≈.88), and when considering only dissensus estimates participants (M=0.56 vs. 0.51, p≈.57).

Figure 3.9. *Mean self-weight in revisions among revisers' responses, Experiment 2*



Top performers revised their estimates with slightly higher self-weights than bottom performers did (M=0.59 vs. 0.54, p≈.52). However when inspecting self-weight patterns by types of initial estimates (Figure 3.9), top performers applied lower self-weights than bottom performers did to consensus initial estimates (M=0.63 vs. 0.69, p≈.98 in the high-variance condition; M=0.44 vs. 0.68, p≈.73 in the low-variance condition), but top performers applied higher self-weights than bottom performers did to dissensus initial estimates (M=0.64 vs. 0.48, p<.10 in the high-variance condition; M=0.58 vs. 0.45, p≈.24 in the low-variance condition). The differences in all cases were not significant.

3.5.2.2 Comparison of revisions in Experiments 1 and 2

Revision levels. Contrary to what expected, participants in this experiment chose self at a higher proportion, although not significantly than participants in the first

experiment (Figure 3.10) whether when initial estimates were consensus (M=0.49 vs. 0.44, p≈.56 in the high-variance condition; M=0.57 vs. 0.52, p≈.82 in the low-variance condition), or dissensus (M=0.29 vs. 0.16, p≈.14 in the high-variance condition; M=0.20 vs. 0.11, p≈.25 in the low-variance condition).

Figure 3.10. *Proportions of choose-self revisions, Experiment 1 & 2*

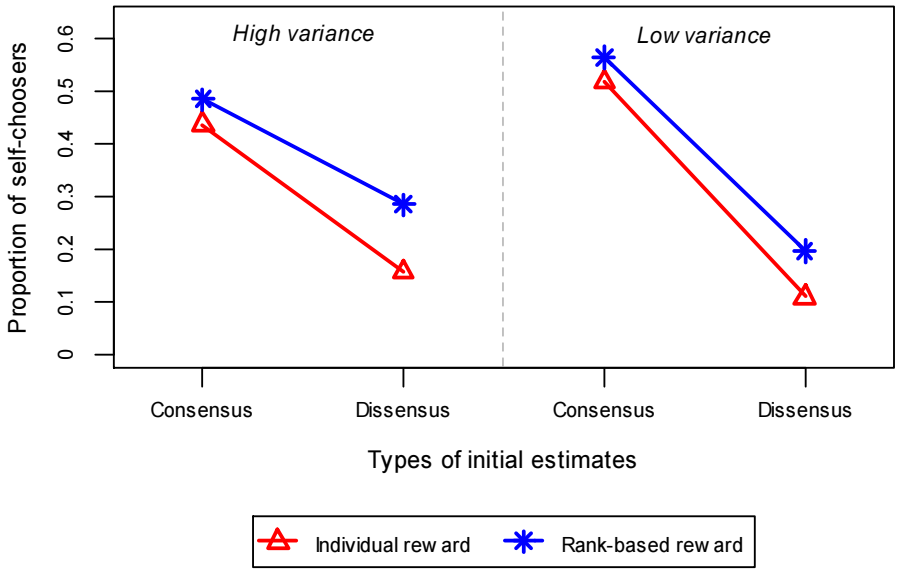
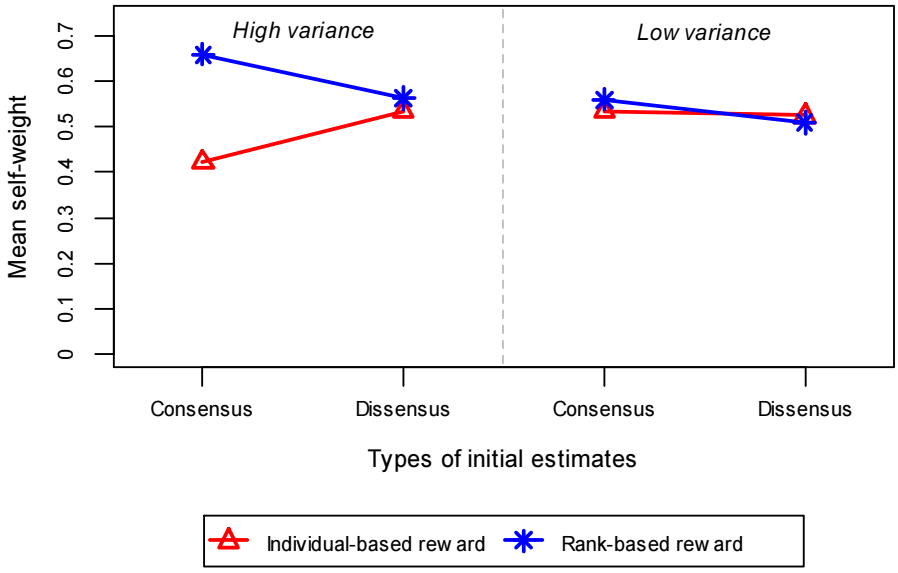


Figure 3.11. *Mean self-weight in revisions among revisers' responses, Experiment 1 & 2*



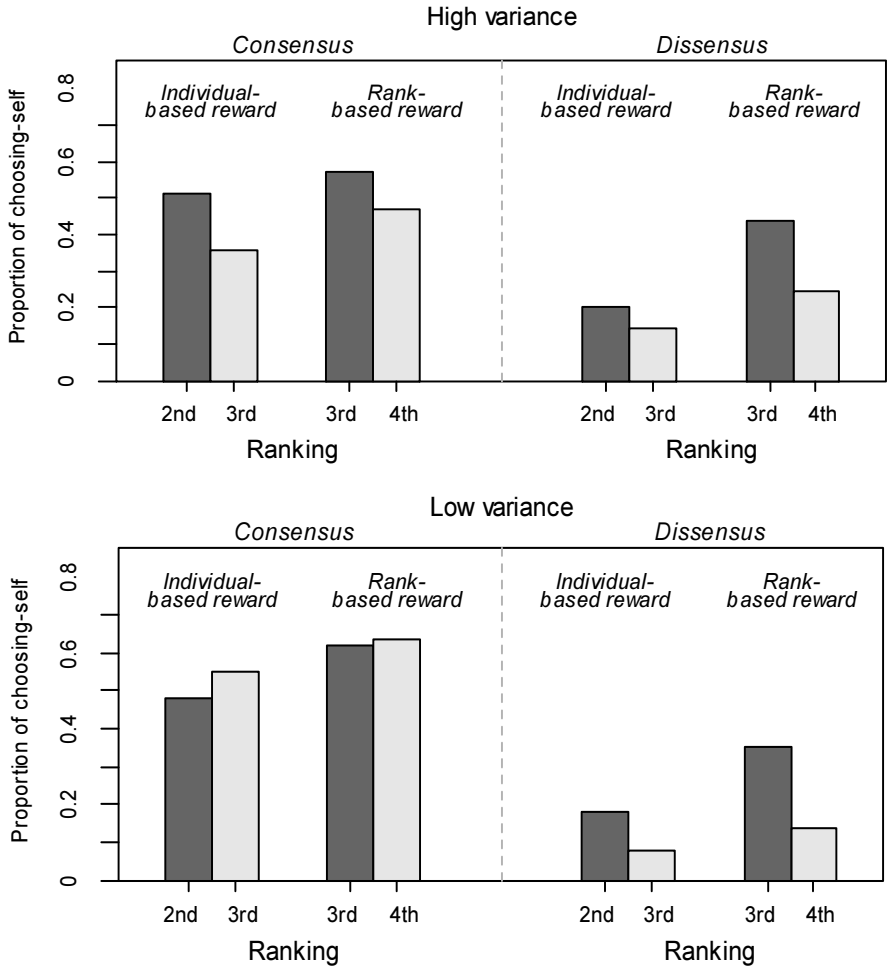
Also, as depicted in Figure 3.11, contrary to what expected this experiment saw significantly higher self-weights applied to consensus initial estimates than the first experiment (M=0.66 vs. 0.42, p<.05 in the high-variance condition; M=0.56 vs. 0.53, p<.05 in the low-variance condition). But with dissensus initial estimates, the differences in self-weights between two experiments were small and not significant (M=0.56 vs. 0.53, p≈.57 in the high-variance condition; M=0.51 vs. 0.52, p≈.83 in the low-variance condition).

Effect of rankings. From the results above, the design of the rank-based reward scheme in this experiment did not produce any significantly higher revision levels than those from Experiment 1 as expected. Reasons could include that participants in this experiment did not take ranking into consideration as intended. But expected ranking depended on how participants perceived the likelihood that their own estimates were the most accurate, in which case then ranking is clearly 1, compared to the likelihood that the SOs were the most accurate, in which case the ranking depends on how many advices were closer to the SO's than their own estimates. Since I did not collect the information on such perceived accuracy likelihoods, only the SO-based ranking was featured in the following analyses. Figure 3.12 shows the proportions of choosing-self under different rankings.

Among consensus estimates (whose rankings could be either 2nd or 3rd), the results from the GEE logit regressions revealed that the difference in the proportions of choosing-self between ranking 2nd or 3rd was not significant in either the first experiment with an individual-based reward scheme (M=0.52 vs. 0.36, p≈.16 in the high-variance condition; M=0.48 vs. 0.55, p≈.22 in the low-variance condition), or the second experiment with an rank-based reward scheme (M=0.54 vs. 0.65, p≈.55 in the high-variance condition; M=0.19 vs. 0.21, p≈.87 in the low-variance condition). This is not unexpected as participants were likely to perceive their consensus estimates to be at least equally accurate compared to the SOs, making the rankings used in the regressions here irrelevant. For dissensus estimates (whose rankings could be either 3rd or 4th) from which participants should expect accuracies inferior to the SOs, making rankings more relevant, indeed in the rank-based second experiment, ranking better at 3rd resulted in lower proportions of choosing-self than ranking worse at 4th as expected, and the difference approached significant in the high-variance condition (M=0.44 vs. 0.25, p<.10), and it was significant in the low-variance condition (M=0.35 vs. 0.14, p<.05). While the same patterns were also found in the first experiment 1, the

difference was not significant (M=0.20 vs. 0.14, p≈.56 in the high-variance condition; M=0.18 vs. 0.08, p≈.34 in the low-variance condition).

Figure 3.12. *Proportions of choose-self revisions by rankings, variance levels, and types of initial estimates, Experiment 1 & 2*



3.5.2.3 Discussions

Largely, the factors on revisions decisions that were tested exerted similar results as in Experiment 1, with the consensus-dissensus category being a significant factor during the first step of revision decision where participants decided whether or not to revise. Beyond that I found no consistently significant pattern of how hypothesized factors affected the decisions in any of the two steps, in fact in certain conditions, some factors showed effects opposite of what expected. When compared to

Experiment 1, the results of this experiment did not exhibit significantly higher levels of advice-taking. However ranking was not totally a non-issue for experimental participants. Further analyses did reveal that there were times, specifically when an initial estimate was a dissensus, that the concern that ranking of participants' initial estimates significantly affected first-stage decisions whether to revise. The pattern between choosing self and rankings was similar in Experiment 1, but it was probably due to the fact that worse-ranking estimates were on average farther from the given advices, making them look more dissenting than better-ranking ones. In the first experiment, ranking per se (or if interpreting the distance from the SO in terms of ranking) was not a significant factor.

That a level of advice taking found in Experiment 2 was not higher than what found in Experiment 1 maybe due to that participants in Experiment 2 did not consider ranking as an important factor, or that being worse ranked did not come with a reward level low enough to trigger herding. Compared to the real-world setting, this experiment's design could not create continuous and cumulative pressure that social comparison usually exerts, such that a current ranking can have implication on future rewards. Moreover, researches in estimators' herding are principally from the field of finance. The high revision level when reputation is of a concern might be due to idiosyncratic nature of the industry that the design of Experiment 2 did not capture. Also, as mentioned earlier, ranking is only a part of what a reputation consists of.

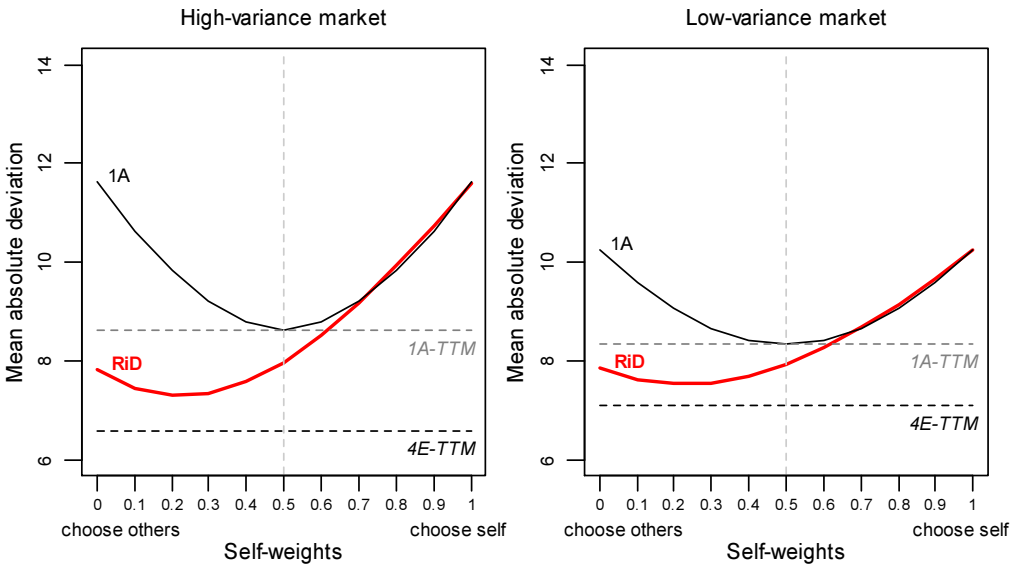
3.6 Benefit of multiple advices: revise-if-dissensus heuristic

Results from Experiment 1 showed that participants decided whether to revise their estimates based on how they appeared in comparison to advices. A consensus estimate was likely to be retained, while a dissensus estimate was likely to receive a revision. In fact, analyses of estimates' accuracies revealed that this categorization was a reliable tool to gauge an expected accuracy of an estimate, that is dissensus estimates were less accurate than advices while consensus estimates were as good as advices, even as good as a recommended revision heuristic such as TTM. This way of categorizing an estimate is only possible when a judge receives more than one advice.

To more completely explore the benefit of having multiple advices I conducted a simulation study using estimates collected from both Experiments 1 and 2, which

totaled to 317 estimates, outliers excluded, from the high-variance market, and 320 estimates from the low-variance market. Estimates for the same house of the same market were grouped together, and within each group I constructed exhaustive permutations of four-estimates sample set. There were a total of 1,004,646 and 1,048,326 sets in the high- and low-variance markets respectively. The first estimate in each set was assigned to be of a judge, the second estimate of the first advisor, the third estimate of the second advisor, and the fourth estimate of the third advisor. Four types of revision strategies were computed for each set. Two assumed that a judge received three advices: 1) revise-if-dissensus heuristic revision, or RiD, where a judge retained her estimate when it is a consensus, and revised her estimate only when it was a dissensus by combining her estimate with the median advice using a certain level self-weight, and 2) take-the-median of all four estimates, or 4E-TTM, where a judge always took the median of the four estimates as her revision. The other two assumed that a judge received only an advice, i.e an estimate from the first advisor: 3) one-advice combination, or 1A, where a judge combined her initial estimate with an advice using a certain level self-weight, and 4) 1-advice take-the-median, or 1A-TTM, where a judge used a middle value between her estimate and an advice. Accuracy of each revision was measured by MAD from the true-model price of the house that each estimate targeted. The results are shown in Figure 3.13.

Figure 3.13. *Mean absolute deviations from true-model prices, by estimate revision strategies*



The benefit of multiple advices is clear. It is obvious that accuracies of TTM improve with the number of estimates in a set, 4E-TTM has lower MADs than 1A-TTM in both levels of variance. However DMs does not often choose the median as a revision. For that, having multiple advices can improve accuracies of revisions especial when there is a high level of market uncertainty. RiD of four advices, itself a sub-optimal strategy when there are multiple advices, outperforms 1A-TTM, that is the best strategy when having only one advice, when DMs choose a self-weight below 0.8, which is a high threshold comparing to the average self-weights below 0.5 that participants in this study decided to use. When the variance is low, RiD yields a similarly good level of accuracy at a self-weight of 0.5 or below which was approximately what experimental participants applied to revise their estimates.

3.7 General discussion

Time and time again, studies have shown that people are likely to adhere to their own initial opinions, and tend to see opinions of others through the subjective lens that their own opinions shape. This results in under-utilization of information that other opinions can contribute. Explanations for such bias include that people are generally over-confident, that they know how their own opinions are formed but not how others are, that they do not want their efforts and other resources that they have spent to arrive at certain opinions to go to waste, or just that people are generally reluctant to change. However, most of the studies on advice-taking concern the case where DMs receive only one other opinion, while it is not unusual that DMs would seek at least a few advices before making final decisions. The main purpose of this chapter is to examine whether people, being social animals that are subject to a peerpressure, would choose utilize advices more to conform to a group consensus.

Results from the first experiment demonstrate that people are much more willing to adjust their opinions, in this case numerical estimates of uncertain quantity, if they appear to be an odd-man-out dissensus. On the other hand people are also more satisfied with opinions if theyare already close to a group's consensus. This fast and frugal (Goldstein & Gigerenzer, 2004) heuristic of revise-if-dissensus appears to be ecologically valid as dissensus opinions indeed prove to be in need of a revision, unlike consensus opinions. The following experiment tests if adding a reward-compatible

pressure of social comparison will encourage a higher degree of advice taking as previous research has proposed. The results of the second experiment also show that people's penchant to conform is translated into opinion revision. However the new ranking-based reward scheme does not reduce further either the tendency to retain original estimates or the level of self-weight used to combine estimates.

Soll and Larrick (2009) argue that once DMs decide to revise, they are non-egocentric taking a simple average between their estimates and an advice they receive. The model used in this chapter assumes that combine their estimates with one single estimate, the SO, that represents all advices that DMs receive. As largely all factors explored here do not have significant effect over the level of self-weights used for estimate combination, so it is possible that in the set-up of this study DMs also simply take the average between their estimates and the SOs. But distributions of self-weights here do not feature a peak around 0.5 like what is exhibited in the study by Soll and Larrick (2009). That the peaks of distributions are located above 0.5 suggests that the SOs skew towards DMs' estimates, which in turn indicates that DMs anchor at their own estimates when choosing a representative of each advice set. In doing so, they exert an egocentric bias indirectly over revision choices. While this bias adversely affects accuracy improvement in revisions, it is not likely to be so large that it nullifies the advantage of multiple advices over a single advice.

One factor unexamined here that future research can extend from this study is the effect of a group size. Groups with different numbers of members exert differentially a pressure to conform (Asch, 1951, Gerard et al., 1968), and one could expect that the larger an advice set, the more likely DMs will revise. However with four advices or more, for example, DMs estimates could be falling into a space between advices but not adjacent to a median of an advice set, making it neither a clear consensus nor a definite dissensus. When that happens, it is possible that DMs, subject to a confirmation bias, will interpret surrounding advices supportive of their own opinions as close to a consensus than they objective are (Gilovich, 1990), and the situation could simply obviate revision. Also, as estimate aggregation is affected by the skew of an estimate set (see Chapter 1), more advices means any single outlying advice will not be so weighty over inference of the SO. More advices might even reduce the degree that egocentricity biases how DMs infer a representative opinion.

References

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting*, 5(4), 585-588.

Armstrong, J. S. (2001). Principles of forecasting: A handbook for researchers and practitioners. Dordrecht, Netherlands: Kluwer.

Asch, S.E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), Groups, leadership and men. Pittsburgh: Carnegie Press.

Ashton H. A. & Ashton R. H (1988). Sequential belief revision in auditing. *The Accounting Review*, 63(4), 623–641.

Ashton, A. H. & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, 31(12), 1499-1508.

Atalay, A. A., Bodur, H. O. & Rasolofoarison, D. (2012). Shining in the center: Central gaze cascade effect on product choice. Journal of Consumer Research, 29(4), 848-866.

Beach, L. R. & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3(3), 439-449.

Beach, L. R. & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance*, 3(2), 109-123.

Beach, L. R. & Solak, F. (1969). Subjective judgments of acceptable error. *Organizational Behavior and Human Performance*, 4(3), 242-251.

Blattberg, R. C. & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36(8), 887-899.

Boldry, J. G., Gaertner, L. & Quinn, J. (2007). Measring the measures: A meta-analytic investigation of the measures of outgroup homogeneity. *Group Processes Intergroup Relations*, 10(2), 157-178.

Bonaccio, S. & Dalal, S. R. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Performance*, 101(2), 127-151.

Brown, L. D., Call, A. C., Clement, M. B., and Sharp, N. Y. (2013). Inside the 'black box'of sell-side financial analysts. Available at SSRN:http://ssrn.com/abstract=2228373.

Budescu, D. V., Rantilla, A. K., Yu, H. & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, 90(1), 178-194.

Bunn, D. W. (1996), Non-traditional methods of forecasting. *European Journal of Operational Research*, 92(3), 528-536.

Carlson, K. A. & Russo, J. E. (2001). Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied*, 7(2), 91–103.

Carlson, K. A., Meloy, M. G. & Russo, J. E. (2001). Leader-driven primacy: Using attribute order to affect consumer choice. *Journal of Consumer Research*, 32(4), 513–518.

Carney, D. R. & Banaji, M. R. (2012). First is best. *Journal PLoS ONE*, 7(6).

Chevalier, J. &Ellison, G. (1999). Career concerns of mutual fund managers. *The Quarterly Journal of* Economics, 114(2), 389-432.

Christensen-Szalanski, J. J. (1978). Problem solving strategies: A selection mechanism, some implications, and some data. *Organizational Behavior and Human Performance*, 22(2), 307–323.

Clarke, K. A. (2003). A simple distribution-free test for nonnested hypotheses. *Political Analysis*, 15(3), 347-363.

Clarke, K. A. (2007). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, 47(1), 72-93.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), 559-583.

Clemen, R. T. & Winkler, R. L. (1986). Combining economic forecasts. *Journal of Business & Economic Statistics*, 4(1), 39-46.

Clement, M. B. & Tse, S. Y. (2005). Financial analyst characteristics of herding behavior in forecasting. *The Journal of Finance*, 40(1), 307-341.

Collett, D. & Lewis, T. (1976). The subjective nature of outlier rejection procedures. *Applied Statistics*, 25(3), 228-237.

Curley, S. P., Young, M. J., Kingry, M. J. &Yates, J. F. (1988). Primacy effects in clinical judgments of contingency. *Medical Decision Making*, 8(3), 216-222.

Dahl, L. C., Brimacombe, C. A. E. & Lindsay, D. S. (2009). Investigating investigators: How presentation order influences participant-investigators' interpretation of eyewitnesses identification and alibi evidence. *Law and Human Behavior*, 33(5), 368-380.

Dawes, R. M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81(2), 95-106.

De Bruin, W. B. & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes*, 93(1-2), 91–101.

de Menezes, L. M. , Bunn, D. W. & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120(1), 190-204.

Doyle, P. & Fenwick, I. A. (1976). Sales forecasting-using a combination of approaches. *Long Range Planning*, 9(1), 60-69.

Einhorn, H. J. & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13(2), 171-192.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance*, 6(1), 1-27.

Einhorn, H. J. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, 59(5), 562-571.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 7(1), 80-106.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2). 117–140.

Fischer, I. & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, 15(3), 227-246.

Flores, B. E. & White, E. M. (1989). Subjective versus objective combining of forecasts: An experiment. *Journal of Forecasting*, 8(3), 331-341.

Frey, D. (1981). Reversible and irreversible decisions: Preference for consonant information as a function of attractiveness of decision alternatives. *Personality and Social Psychology Bulletin*, 7(4), 621-626.

Froot, K. A., Scharfstein, D. S. & Stein, J. C.(1992). Herd on the street: Informational inefficiencies in a market with short-term Speculation. *Journal of Finance*, 47(4), 1461-1484.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *The Journal of General Psychology*, 113(4), 351–357.

Galton, F. (1907a). One vote, one value. *Nature* 75, 414. http://galton.org/cgi-bin/searchImages/galton/search/essays/pages/galton-1907-vote-value.htm.

Galton, F. (1907b). Vox populi. *Nature* 75, 450-51. http://galton.org/cgi-bin/searchImages/galton/search/essays/pages/galton-1907-vox-populi_1.htm.

Gerard, H. B., Wilhemy, R. A. & Conolle, E. S. (1968). Conformity and group size. *Journal of Personality and Social Psychology*, 8(1), 79-82.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4), 650-669.

Gilliland, S. W., Schmitt, N. & Wood, L. (1993). Cost-benefit determinants of decision process and accuracy. *Organizational Behavior and Human Decision Processes*, 56(2), 308–330.

Gilovich, T. (1990). Differential construal and the false consensus effect. *Journal of Personality and Social Psychology*, 59(4), 623-634.

Goodwin, P. & Wright G. (1994). Heuristics, biased and improvement strategies in judgmental time series forecasting. Omega, *International Journal of Management Science*, 22(6), 553-568.

Greenwald, A. G. (1980). Fabrication and revision of personal history. *American Psychologist*, 35 (7), 603-618.

Guiral-Contreras, A., Gonzalo-Angulo, J. A. & Rodgers, W. (2007). Information content and recency effect of the audit report in loan rating decisions. *Accounting and Finance*, 47(2), 285– 304.

Harries, C., Yaniv, I. & Harvey, N. (2004). Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, 17(5), 333-348.

Harris, D. M. & Guten, S. (1979). Health-protective behavior: an exploratory study. *Journal of Health and Social Behavior*, 20 (1),17-29.

Harvey, N. & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70(2), 117-133.

Harvey, N. & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting*, 20(3), 391-409.

Harvey, N., Harries, C. & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes*, 81(2), 252-273.

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

Hausmann, D. & Läge, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, 3(3), 229-243.

Helson. H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review*, 55(6), 297-313.

Hogarth, R. M. & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Science*, 27(2), 115-138.

Hogarth, R. M. (1978). A note on aggregation opinions. *Organizational Behavior and Human Performance*, 21(1), 40-46.

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

Hong, H., Kubik, J. D. & Solomon, A. (2000). Security analysts' career concerns and herding of earnings forecasts. *RAND Journal of Economics*, 31(1), 121-144.

Hulland, J. S. & Kleinmuntz, D. N. (1994). Factors influencing the use of internal summary evaluations versus external information in choice, *Journal of Behavioural Decision Making*, 7(2), 79–102.

Kardes, F. R. & Herr, P. M. (1990). Order effects in consumer judgment, choice, and memory: The role of initial processing goals. *Advances in Consumer Research*, 17(1), 541–546.

Kerstholt, J. H. & Jackson, J. L. (1999). Judicial decision making: order of evidence presentation and availability of background information. *Applied Cognitive Psychology*, 12(5), 445–454.

Lamont, O. (1995). Macroeconomic forecasts and microeconomic forecasters. *Journal of Economic Behavior & Organization*, 48(3), 265-280.

Larrick, R. P. & Soll, J. B. (2006). Intuition about combining opinions: Misappreciation of the averaging principle. *Management Science*, 52(1), 111-127.

Lathrop, R. G. (1967). Perceived variability. *Journal of Experimental Psychology*, 73(4), 498-502.

Lawrence, M. & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, 43(2), 172-187.

Lawrence, M., Edmundson, R. H. & O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science*, 32(12), 1521-1532.

Lawrence, M., Goodwin, P., O'Connor, M. & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3), 493-518.

Lees, C. D. & Triggs, T. J. (1997). Intuitive prediction: Response strategies underlying cue weights in the relative-weight averaging model. *American Journal of Psychology*, 110(3), 317-356.

Levin, I. P. (1976). Processing of deviant information in inference and descriptive tasks with simultaneous and serial presentation. *Organizational Behavior and Human Decision Performance*, 15(2), 195–211.

Libby, R. & Blashfield, R. K. (1978). Performance of composite as a function of the number of judges. *Organizational Behavior and Human Performance*, 21(2), 121-129.

Lim, J.S. & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: its effectiveness and biases. Journal *of Behavioral Decision Making*, 8(3), 149-168.

Maines, L. A. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting*, 12(2), 223-233.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2), 111-153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K. & Simmons, L. R. (1993). The M-2 competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9(1), 5-22.

Mantel, S. P. & Kardes, F. R. (1999). The role of direction of comparison, attribute-based processing, and attitude-based processing in consumer preference. *Journal of Consumer Research*, 25(3), 335-52.

Mantonakis, A., Rodero, P., Lesschaeve, I. & Hastie, R. (2009). Order in choice. Effects of serial position on preferences. *Psychological Science*, 20(11), 1309–1312.

Marsh J. K. & Ahn, W. K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, 34(3), 568–576.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166.

Minson, J. A. & Mueller, J. S. (2012). The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science*, 20(10), 1-6.

Newbold, P. & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. Jo*urnal of The Royal Statistical Society: Serie A*, 137(2), 137-146.

O'Reilly, D. M., Leitch, R. A. & Wedell, D. H. (2004). The effects of immediate context on auditors' judgments of loan quality. *Auditing: A Journal of Practice & Theory*, 23(1), 89–105.

Patt, A. G., Bowles, H. R. & Cash, D. (2006). Mechanisms for enhancing the credibility of an advisor: prepayment and aligned incentives. *Journal of Behavioral Decision Making*, 19(4), 347-359.

Payne, J. W., James R. Bettman, J. R. & Johnson, E. J. (1988). Adaptive strategy selection in decision making. J*ournal of Experimental Psychology: Learning, Memory & Cognition*, 14(3), 534-52.

Payne, J. W., Bettman, J. R. & Johnson, E. J. (1993). The Adaptive Decision-Maker. Cambridge University Press, 1993.

Perloff, L. S. & Fetzer, B. K. (1986). Self–other judgments and perceived vulnerability to victimization. *Journal of Personality and Social Psychology*, 50(3), 502-510.

Peterson, C. R. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68(1), 29-46.

Rantilla, A. K. & Budescu, D. V. (1999). Aggregation of expert opinions. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

Reneau, J. H. & Blanthorne, C. (2001). Effects of information sequence and irrelevant distractor information when using a computer-based decision aid. *Decision Sciences*, 32(1), 145–163.

Ritov, I. & Baron, J. (1992). Status-quo and omission bias. *Journal of Risk and Uncertainty*, 5(1), 49-61.

Russo, J. E., Meloy, M. G. & Medvec, V. H. (1998), Predecisional distortion of product information. *Journal of Marketing Research*, 35(4), 438-452.

Saad, G. & Russo, J. E. (1996). Stopping criteria in sequential choice. *Organizational Behavior and Human Decision Processes*, 67(3), 258–270.

Samuelson, W. & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1(1), 7-59.

Scharfstein, D. S. & Stein, J. C.(2001). Herd behavior and investment. *The American Economic Review*, 80(3), 465-479.

Sniezek, J. A. & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, 43, 1-28.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, 38(2), 317-346.

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

Todd, P. & Benbasat, I. (1994). The influence of decision aids on choice strategies: An experimental analysis of the role of cognitive effort. *Organizational Behavior and Human Decision Processes*, 60(1), 36-74.

Trueman, B. (1994). Analyst forecasts and herding behavior. *Review of Financial Studies*, 7(1), 97-124.

Tubbs, R. M., Gaeth, G. J., Levin, I. P. & van Osdol, L. A. (1993). Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making*, 6(4), 257–269.

Webby, R. & O'Connor, M. (1996). Judgmental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, 12(1), 91-118.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5), 806–820.

Winkler, R. L. & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operation Research*, 40(3), 609-614.

Winkler, R. L. & Clemen, R. T. (2004). Multiple experts vs. multiple methods: Combining correlation assessments. *Decision Analysis*, 1(3), 167-176.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*. 66(4), 675-685.

Xu, Y. & Kim, H. W. (2008). Order effect and vendor inspection in online comparison shopping. *Journal of Retailing*, 84(4), 477–486.

Yaniv, I. & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, 124(4), 424-432.

Yaniv, I. & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information Asymmetry and combination rules. *Psychological Science*, 4(1), 58-62.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69(3), 237-249.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. Organizational *Behavior and Human Decision Processes*, 93(1), 1–13.

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, 13(2), 76–79.

Yaniv, I. & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, 10(1), 21–32.

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

Yaniv, I. & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness tradeoff. *Journal of Experimental Psychology: General*, 124(4), 424–432.

Yaniv, I. & Milyavsky, M. (2007). Using advice from multiple sources torevise and improve judgment. *Organizational Behavior and Human Decision Processes*, 103(1), 104 –120.

Yaniv. I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69(3), 237-249.

Yates, F. M. & Curley, S. P. (1986). Contingency judgment: Primacy effects and attention decrement. *Acta Psychologica*, 62(3), 293–302.