MANAGEMENT SCIENCE
Vol. 52, No. 1, January 2006, pp. 111–127
ISSN 0025-1909 | EISSN 1526-5501 | 06 | 5201 | 0111

informs

DOI 10.1287/mnsc.1050.0459

Intuitions About Combining Opinions: Misappreciation of the Averaging Principle

Richard P. Larrick, Jack B. Soll
Fuqua School of Business, Duke University, Durham, North Carolina 27708 {larrick@duke.edu, jsoll@duke.edu}

Averaging estimates is an effective way to improve accuracy when combining expert judgments, integrating group members' judgments, or using advice to modify personal judgments. If the estimates of two judges ever fall on different sides of the truth, which we term bracketing, averaging must outperform the average judge for convex loss functions, such as mean absolute deviation (MAD). We hypothesized that people often hold incorrect beliefs about averaging, falsely concluding that the average of two judges' estimates would be no more accurate than the average judge. The experiments confirmed that this misconception was common across a range of tasks that involved reasoning from summary data (Experiment 1), from specific instances (Experiment 2), and conceptually (Experiment 3). However, this misconception decreased as observed or assumed bracketing rate increased (all three studies) and when bracketing was made more transparent (Experiment 2). Experiment 4 showed that flawed inferential rules and poor extensional reasoning abilities contributed to the misconception. We conclude by describing how people may face few opportunities to learn the benefits of averaging and how misappreciating averaging contributes to poor intuitive strategies for combining estimates.

Key words: averaging opinions; combining forecasts; information aggregation; advice taking; heuristics and biases
History: Accepted by Detlof von Winterfeldt, decision analysis; received August 27, 2003. This paper was with the authors 11 months for 2 revisions.

Introduction

An old saying has it that two heads are better than one. This aphorism first gained scientific support in the 1920s and 1930s when psychologists discovered that averaging individual quantity judgments led to more accurate estimates than those of the average individual judge (Bruce 1935, Gordon 1924, 1935, Knight 1921, Smith 1931). However, a debate ensued. Why did so-called "statisticized" groups outperform the average individual? Is there truly something special about groups? As one writer put it, "In every coming together of minds . . . [t]here is the Creative Plus, which no one mind by itself could achieve" (Overstreet 1925, cited in Watson 1928, p. 328). The ultimate conclusion, however, was that statisticized groups revealed nothing special about groups of people, but confirmed instead the statistical principle that aggregation of imperfect estimates reduces error (Eysenck 1939, Kelley 1925, Preston 1938, Stroop 1932). In hindsight, it is surprising that it took two decades of theoretical arguments and empirical demonstrations to accept this conclusion (Lorge et al. 1958, Steiner 1972).

Why did it take so long to grasp such a simple statistical principle? The answer, we believe, is that the principle is not so simple to recognize. People lack the intuition for it, and rarely have an opportunity to learn it. A consequence of misunderstanding averaging is that people use inferior approaches to aggregating uncertain quantity estimates. Individuals inappropriately use advice from others, frequently ignoring it or occasionally accepting it completely (Soll and Larrick 2005). Judges avoid combining estimates across sources, such as forecast models that make different assumptions, because they do not understand that averaging reduces error (Soll 1999).

This paper focuses on how well people recognize what we call the averaging principle: When combining uncertain quantity estimates, the discrepancy between the average estimate and the truth can be no greater than the average discrepancy of the component estimates. To illustrate the principle intuitively, imagine two people forecasting the high temperature in Honolulu tomorrow, which turns out to be 73°. If they guess 60° and 70°, they miss by 13° and 3°, respectively, or 8° on average. The average guess, 65°, also misses by 8°. Here, the average estimate performs equally well as the average judge. Now imagine they guess 60° and 80°, so that the two estimates "bracket" the truth. In this instance, their guesses miss by 13° and 7°, or 10° on average. But the average guess of 70° misses by only 3°. Averaging outperforms the average individual (and, in this case, happens to outperform both individuals).

An important implication of the averaging principle is that, over multiple judgments, the mean absolute deviation (MAD) of averaging is less than the MAD of the average individual if there is at least one instance of bracketing.[1] We conceive of deviation from the truth (which we will shorten to "error") as having two components: *Systematic bias*, which is defined as a signed constant for a given judge over a given set of judgments, and *random error*. As the actual rate of bracketing increases, so does the power of averaging. For example, two judges who have normally distributed random errors that are unbiased (i.e., mean-zero) and uncorrelated will bracket the truth 50% of the time. If the judges have approximately similar skill (that is, they have similar MADs), averaging will improve accuracy by about 29%. The bracketing rate will be lower than 50% if the two judges share a bias (e.g., across a set of cities both judges tend to overestimate or underestimate the truth), or if they have positively correlated random errors (e.g., underestimating in winter and overestimating in summer), in which case averaging will improve accuracy by less than 29%. Analogously, opposing biases and negatively correlated random errors will result in bracketing rates greater than 50%, in which case averaging will improve accuracy by more than 29%. The critical conclusion is that, as long as there is any bracketing whatsoever, averaging must be more accurate than the average judge.

Extensive research in the forecasting (Armstrong 2001, Clemen 1989), decision making (Ariely et al. 2000, Johnson et al. 2001, Wallsten et al. 1997), and groups (Einhorn et al. 1977, Gigone and Hastie 1997) literatures has confirmed that averaging is a powerful and robust way of reducing error in quantitative judgment. Similar conclusions have been reached for other aggregation mechanisms, such as simple majority rule (Hastie and Kameda 2005, Sorkin et al. 1998). Although applied statisticians originally treated averaging as a baseline against which to compare more sophisticated combination methods, the baseline proved surprisingly difficult to beat (Clemen and Winkler 1986, Fildes and Makridakis 1995). In an extensive review, Armstrong (2001) reanalyzed 30 studies conducted between 1960 and 2000. Across the studies, averaging improved forecast accuracy from 3.4% to 23.5% relative to the mean performance of the forecasts being averaged, with a mean improvement of 12.5% (see Surowiecki 2004 for an engaging overview of the benefits of aggregation and when they can be realized).

Averaging itself is not critical to improving judgments. Weights in rough proximity to whatever weights are analytically optimal yield similar levels of improvement (Dawes and Corrigan 1974, von Winterfeldt and Edwards 1986). For example, if the appropriate weighting scheme is 0.5/0.5, using a 0.7/0.3 split will be nearly as accurate. However, empirical studies indicate that people frequently do not combine at all, but instead choose between estimates (Soll and Larrick 2005). In one of our studies, participants first estimated salaries for graduates of 25 business schools. They were paid based on the MAD of their revised estimates after viewing the responses of another participant. Although participants reduced MAD by 10% with their intuitive revision strategies (see also Harvey and Fischer 1997, Yaniv 2004), they would have improved by 16% had they consistently averaged. Because participants used extreme weighting schemes of 1/0 or 0/1 about half the time (strategies we call "choosing"), they often missed out on the benefits of combination.

Why People Misappreciate Averaging

Although many factors may contribute to the misappreciation of averaging, we will focus on people's inability to reason extensionally (Kahneman and Frederick 2002, Tversky and Kahneman 1983, Stanovich and West 2000) and their tendency to rely instead on flawed inferential rules and aphorisms (Holland et al. 1986, Nisbett and Ross 1980). Extensional reasoning involves partitioning events into subsets that are mutually exclusive and exhaustive, and then aggregating over these subsets to draw inferences (cf. Tversky and Kahneman 1983). In the case of averaging, two judges can either bracket the truth or not. When the estimates bracket, averaging performs better than the average judge; when the estimates do not bracket, averaging performs equally well as the average judge. Putting the two cases together, averaging can do no worse than the average judge, and will do better across a set of questions provided that there is at least one instance of bracketing.

Where might people go wrong in this sequence? If they do attempt to reason extensionally, they may sample possible instances incompletely, especially if they rely on their own imagination (Fiedler 2000). Extensional reasoning in itself does not guarantee a correct answer. For example, a person may simulate an instance of no bracketing, stop searching, and falsely conclude that averaging will lead to an average level of accuracy. Alternatively, if a judge simulates an instance of bracketing and stops, the judge

[1] The basic principle that averaging outperforms the average judge holds for any convex loss function. If squared deviation is used, for example, averaging is even more powerful because it outperforms the average individual even if both estimates err in the same direction. Averaging is less attractive for loss functions that are not convex everywhere, such as the log of the deviation. Even for such functions, however, averaging will perform well at high levels of bracketing.

will be quite optimistic about the prospects of averaging. Indeed, in Experiment 3 we find that while many people are pessimistic about averaging, some are overly optimistic, and this is linked to their beliefs about bracketing.

It is also possible that, instead of generating possible instances to assess the effectiveness of averaging, people rely on the representativeness heuristic. One manifestation of this heuristic is that people tend to anticipate outcomes that are representative of the process that generates them. For example, given 20 rolls of a six-sided die with four green faces (G) and two red faces (R), the sequence RGRRR is more likely to appear somewhere in 20 rolls than will GRGRRR, because the former sequence is included in the latter. However, most people prefer to bet on the longer, less likely sequence (Tversky and Kahneman 1983). An explanation for this result is that people assess the similarity of the outcome to what they expect from the underlying process (i.e., rolling a die with more green than red sides), and then substitute this judgment of similarity to form a judgment of probability (Kahneman and Frederick 2002).

A similar explanation might account for how people assess the performance of averaging quantity estimates. Average performance is more similar to the process of averaging than is accurate performance, both semantically, and in the sense that both involve computing an arithmetic mean. If people substitute this similarity judgment when they evaluate the performance of averaging, they will conclude that averaging leads to an average result. In the case of averaging the opinions of an expert and a novice, the representativeness heuristic would lead people to conclude that the expert's performance will be dragged down and the novice's will be lifted up. In the case of two novices, they would conclude that performance is locked in at this mediocre level. Such representativeness-based reasoning is captured in cultural aphorisms (Nisbett and Ross 1980) such as "you can't get something from nothing" or "gigo—garbage in, garbage out." Other cultural rules may also compete with or substitute for extensional reasoning, which we return to in the introduction to Experiment 4.

Despite such failures in extensional reasoning, better reasoning is facilitated when information environments are more transparent (Brase 2002, Hoffrage et al. 2000, Lewis and Keren 1999, Nisbett et al. 1983). This suggests that averaging will be easier to appreciate when bracketing is easier to observe and encode. Consider two judges who are forecasting tomorrow's high temperatures for 10 U.S. cities. When the estimates of the two judges are observed simultaneously *and* with the true outcome, it will be apparent to an observer that the truth often does "lie in the middle." In such an environment, people will also be sensitive to the bracketing rate between two judges, predicting more benefit from averaging as the bracketing rate increases. In the absence of simultaneous observation, however, bracketing is essentially undetectable. If judge 1's predictions for Chicago, St. Louis, and so on are observed with the truth, *followed* by judge 2's predictions for the same cities, it will be difficult for an observer to remember specific predictions well enough to recognize bracketing. In this sequential observation environment, people must rely on imagining the space of possible instances or substitute other judgments to predict the effect of averaging. In such an environment, people may never realize that bracketing occurs, and they may never reach the insight that it matters.

The failure to encode bracketing during sequential observation has one notable exception: Bracketing that arises from opposing biases will be transparent. If an observer sees that a judge tends to give temperature estimates that are systematically high (e.g., 5° above the truth over a series of estimates), the observer can encode the bias at the judge level. When a judge who is known to overestimate is paired with a judge known to underestimate, an observer can easily imagine the high rates of bracketing that will follow. Thus, we expect that averaging will be appreciated when two judges have opposing biases regardless of whether their judgments are observed sequentially or simultaneously.

Overview of Experiments

This paper presents four experiments that examine people's beliefs about the effect of averaging judgments on reducing error. Although many questions could be asked regarding people's intuitions about aggregation, we wanted to test understanding of a single, fundamental principle: In the presence of bracketing, the average of individual judgments must be more accurate than the average individual judge. People often neglect more sophisticated, and arguably, more difficult aggregation principles (Gonzalez 1994, Soll 1999). Is this basic principle in the lay repertoire for reasoning about aggregation?

In the first three studies, three complementary methods were used to test whether misconceptions about averaging are robust to variations in stimuli, presentation format, and elicitation methods. The first experiment tested people's understanding of the effect of averaging when presented with summary data about judges' estimates. The second experiment investigated whether people could induce the benefits of averaging by observing judges' estimates directly. The third experiment presented minimal numerical information

to test for people's conceptual understanding of the effects of averaging.

We expected that the averaging principle would be difficult to grasp (the dozen or so empirical papers on averaging between the 1920s and 1940s are a testament to this difficulty!) due to the challenges of extensional reasoning and the availability of competing, flawed inferential rules. We have argued, however, that the benefits of averaging are easier to recognize as the bracketing rate increases and as it becomes more transparent. In these studies, we either manipulated (Experiments 1 and 2) or measured (Experiment 3) the bracketing rate—that is, the relative frequency with which the estimates of two judges fall on opposite sides of the truth. Averaging is more effective to the extent that the bracketing rate is high. The bracketing rate itself becomes more transparent when judgments are observed simultaneously or when judges have opposing biases. We expected that increased transparency would facilitate recognition of the averaging principle. A final study (Experiment 4) tested the contribution of extensional reasoning ability and inferential rules to appreciation of the averaging principle.

Experiment 1

Do people hold the principle that averaging must outperform the average judge if there is some degree of bracketing? Experiment 1 tested this question by providing participants with performance summaries for two hypothetical judges. A difference in the accuracy of the judges was created by setting their MADs at different levels (approximately a 12% difference). The bracketing rate for the two judges was manipulated by varying their historical correlation in random error across three levels: Positive correlation (24% bracketing), weak negative correlation (58%), and strong negative correlation (90%). A fourth condition included judges with opposing biases (90% bracketing). We expected that people would increasingly recognize the effectiveness of averaging as the bracketing rate increased. However, we predicted that many participants would equate the accuracy of averaging judgments with the accuracy of the average individual judge.

Method

Participants. Participants were 145 MBA students enrolled in a statistics course at INSEAD. The population is mathematically sophisticated; the median score on the quantitative section of the GMAT was in the 94th percentile.

Materials. Participants read the following scenario:

> Ms. A and Ms. B are currency analysts at two banks. On the first of every month they have the task of forecasting the yen to dollar exchange rate for the following month. The banks use these forecasts in deciding their currency positions. The two banks have decided to merge and now must decide how to make best use of Ms. A and Ms. B. To help with this decision, the new combined bank has analyzed the past forecasts of Ms. A and Ms. B for the 50 months prior to the merger. There is no evidence that the accuracy level for the forecasters has been changing over this period.

The concept of MAD was explained, using specific illustrations, after which the MADs of Ms. A (4.7) and Ms. B (5.3) were presented.

Participants were told that the banks also tracked the frequency with which each forecaster over or underestimated the true exchange rate for the previous 50 months. Participants then saw one of the joint patterns of forecasters' errors in Figure 1 presented as a 2-by-2 table. The bracketing rate increases across the four panels (24%, 58%, 90%, and 90%, respectively). Regardless of how errors are distributed, the presence of bracketing dictates that averaging *must* outperform the average judge in all cases.

The question that elicited the main dependent variable—participants' estimates of the accuracy of the averaging strategy—was embedded in a larger set of questions. Participants were told "Bank officials

Figure 1 Four Conditions of Error Patterns Between Ms. A and Ms. B in Experiment 1

		Ms. B	
Relationship	Ms. A	Overestimate	Underestimate
Strong positive correlation condition	Overestimate	18	4
	Underestimate	8	20
Weak negative correlation condition	Overestimate	10	13
	Underestimate	16	11
Strong negative correlation condition	Overestimate	3	23
	Underestimate	22	2
Opposing biases condition	Overestimate	2	41
	Underestimate	4	3

Notes. Participants saw error patterns for only one condition.

Table 1 Estimates of the Effectiveness of Averaging Judgments by Dyadic Error Pattern (Experiment 1)

Error pattern (Bracketing rate)	Proportion estimating that averaging performs...		MAD estimates			
	No better than the avg. judge	Better than both judges	Median	Mean	Imputed	*n*
Strong pos. *r* (24%)	0.74	0.26	5.00	4.82	4.63	35
Weak neg. *r* (58%)	0.55	0.40	5.00	4.35	3.06	38
Strong neg. *r* (90%)	0.47	0.53	3.77	3.09	0.85	30
Opposing bias (90%)	0.53	0.45	5.00	3.53	0.97	42

Note. Bracketing rate for each condition is given in parentheses. The proportions in columns 1 and 2 do not sum to 1.0; the difference is the proportion of participants giving estimates that fell in the category "Better than the average judge but worse than Ms. A." The imputed value is an estimate of the correct answer based on the assumptions described in footnote 2.

are discussing the following strategies for best using Ms. A and Ms. B:"

Strategy 1. Retain Ms. A as the dollar/euro forecaster, and reassign Ms. B. (*Ms. A alone*)

Strategy 2. Use Ms. A's forecast 60% of the time and Ms. B's forecast 40% of the time. (*Alternating*)

Strategy 3. Average the two forecasts, and use this average as the bank's forecast. (*Averaging*)

Strategy 4. Ask Ms. A and Ms. B how confident they are for each forecast. Use the one who is more confident. If they're tied on confidence, go with Ms. A. (*Confidence*)

Strategy 5. Have Ms. A and Ms. B sit down and discuss their opinions. Require them to agree on a single forecast. (*Discussion*) (Parenthetical labels were omitted in the stimuli.)

Participants were told to assume that Ms. A and Ms. B would continue to perform at their historic levels of accuracy, and to estimate the MAD that the new bank would achieve in its forecasts for the next 50 months if they used each strategy.

Results and Discussion

Participants' estimates for the averaging strategy were coded as *no better than the average judge* if their estimate was equal to 5 or greater and as *better than both judges* if their estimate was less than 4.7, with the remaining estimates coded in an intermediate category (where 5 is the average of Ms. A's and Ms. B's MADs of 4.7 and 5.3). Across all conditions, 57% of participants expected that averaging would perform no better than the average judge. Of these, nearly all (95%) estimated that averaging would perform exactly equal to the average judge's MAD of 5 (which was both the median and modal response).

As shown in Table 1, participants were less likely to misunderstand the effect of averaging as the bracketing rate increased across the four conditions (Kendall's tau-b $= -0.17$, $p < 0.05$, combining the two 90% conditions). However, even when the bracketing rate was 90%, half the participants estimated that averaging would perform no better than the average judge.

Participants' mean estimates of the MADs for averaging showed a similar pattern as the proportions. As may be seen in Table 1, estimates declined as imputed values (our estimate of a "correct" response based on the assumption that the judges had normally-distributed errors)[2] declined across the four conditions. However, in all four conditions, a significant proportion of participants gave estimates for averaging that were higher than the imputed value (proportions $= 0.74, 0.87, 0.80, 0.86$, ps < 0.01 by a binomial test). Thus, although participants were sensitive to conditions that made averaging more effective, they consistently underestimated the magnitude of the benefit.

There was an interesting discrepancy between the mean and the median, which reveals a second pattern. Although 57% of the participants estimated that averaging would perform no better than the average judge, the 43% who estimated that it would perform better than the average judge expected it to perform *substantially* better. Of this 43% minority, 95% correctly expected averaging to outperform *both* judges in these circumstances. Surprisingly, almost no one gave an intermediate estimate between the average judge (5) and Ms. A (4.7). Taking all conditions together, a narrow majority failed to recognize the power of averaging, whereas a substantial minority recognized that in these cases averaging would surpass both judges.

Table 2 presents the medians for averaging and the other four strategies. In contrast to averaging, participants were substantially more accurate in predicting the consequence of *alternating* between Ms. A and Ms. B and of using *Ms. A's forecasts alone*. A priori, the expected values of these two strategies are 4.94 and 4.7 (in all conditions), which were precisely the median values that participants gave in all conditions (see Table 2). It is worth noting that 80% and

[2] To derive the imputed estimates, we first specified the bivariate normal distribution for the judgments of Ms. A and Ms. B that yields, in expectation, the cell frequencies in the 2×2s. It is then straightforward to compute the MAD for averaging.

Table 2 Median Estimates for Five Combination Strategies (Experiment 1)

	Strategy				
Pattern of error (Bracketing rate)	Averaging	Alternating 60/40	Ms. A alone	Most confident	Discussion
Strong pos. *r* (24%)	5.00	4.94	4.70	4.70	4.90
Weak neg. *r* (58%)	5.00	4.94	4.70	4.80	4.70
Strong neg. *r* (90%)	3.77	4.94	4.70	4.00	5.00
Opposing bias (90%)	5.00	4.94	4.70	4.50	4.85
Overall	5.00	4.94	4.70	4.70	4.92

100% of the participants estimated that alternating and using Ms. A's forecast, respectively, would perform better than the average judge. These proportions stand in contrast to the 43% who believed that averaging would perform that well. Table 2 also shows that the median participant estimated that the *confidence* and *discussion* strategies would outperform averaging.

The Table 2 results indicate that participants' mistaken estimates for averaging were not the result of the lazy or careless use of "5" as a focal answer—for all other strategies, the median participant gave a number other than 5. In particular, the estimates for the alternating strategy indicate that participants processed information with effort and care because the correct response requires precise calculation. Many participants were not able, however, to predict the effect of averaging.

By presenting participants with summary data on individual accuracy and dyadic bracketing rates, Experiment 1 tested whether they held and could apply the abstract principle that averaging outperforms the average judge in the presence of bracketing. The results indicated that the majority of participants did not spontaneously reason using this abstract principle, and reasoned instead that averaging performs at the level of the average judge. Experiment 2 was designed to test whether the averaging principle could be applied more accurately with specific, concrete instances.

Experiment 2

Experiment 1 showed that, for summary data, most people failed to apply the abstract principle that averaging outperforms the average judge. As we proposed in the introduction, we expected that people would engage in more accurate extensional reasoning if they could observe specific instances in which two judgments bracketed the truth. Experiment 2 tested whether people could induce the averaging principle from direct experience with sets of judgments that exhibited bracketing.

As we proposed in the introduction, however, we expected that the ability to recognize the effect of averaging would depend on an important environmental variable: Are judges' estimates observed together (as they were in summary form in Experiment 1) or in isolation? Environment is critical because bracketing—which gives averaging its power—is an inherently *dyadic* property. Social psychologists have argued that people tend to encode behavior in terms of stable properties of individuals but overlook situational variables (Ross and Nisbett 1991). Bracketing is a situational variable that is nearly impossible to recognize if observers are attending to individuals in isolation of each other. It is only apparent when multiple estimates of two judges can be compared to the truth *simultaneously*.

As in Experiment 1, we varied the rate of bracketing by creating a no correlation condition, a negative correlation condition, and an opposing biases condition. In addition, we included two formats in Experiment 2, one in which participants saw the estimates of two judges sequentially (10 guesses by one judge, followed by 10 guesses by a second judge), and one in which participants saw the estimates simultaneously. We expected that participants would recognize the benefits of averaging in the simultaneous format, where bracketing is transparent, and that they would be increasingly sensitive as bracketing increased (as in Experiment 1). However, in the sequential format, we expected that the nontransparency of bracketing would lead participants both to misappreciate the effect of averaging (expecting averaging to perform no better than the average judge) and to be insensitive to bracketing rate (expecting averaging to perform the same when random errors are uncorrelated or negatively correlated).

To the extent that people do attend to and encode behavior at an individual level (Ross and Nisbett 1991), we expected one special case in which averaging would be appreciated regardless of learning environment: The case of opposing biases. The tendency for a judge to err systematically high or low in a domain is easily encoded as a stable individual property ("he tends to underestimate attendance" or "she tends to be an optimist about sales"). Consequently, even when judges are observed in isolation, bracketing that is due to opposing biases is readily apparent ("an optimist and a pessimist will tend to offset each other's excesses"). We therefore expected the effect of averaging opposing biases to be recognized regardless of whether judgments were observed sequentially or simultaneously.

Method

Participants. Participants were 263 MBA students enrolled in a statistics course at INSEAD, in a different cohort than those who participated in Experiment 1.

Figure 2 Stimuli Used in the Different Error Conditions in Experiment 2 (Depicted Here as in the Simultaneous Format Condition)

		No correlation condition (Bracketing rate = 40%)		Negative correlation condition (Bracketing rate = 80%)		Opposing biases condition (Bracketing rate = 80%)	
Day	True attendance	Ty's forecast	Chris's forecast	Ty's forecast	Chris's forecast	Ty's forecast	Chris's forecast
1	285	240	230	300	280	260	310
2	315	390	320	310	335	265	350
3	424	440	405	375	480	390	410
4	254	225	245	225	290	265	300
5	176	180	215	160	140	170	195
6	381	430	345	455	390	365	475
7	346	375	435	375	255	285	350
8	103	85	125	110	45	75	165
9	497	490	405	425	585	420	570
10	219	145	275	265	200	175	270

Notes. Participants saw data (day, true attendance, and forecasts) for only one condition.

Materials. Participants were presented with a scenario about two managers, Ty and Chris, who co-manage a small movie theater. They were told:

> Every morning they predict the attendance at the theatre for that evening. They use the forecast to decide how many employees are needed to staff the theatre. If they underestimate the true attendance the theatre loses revenue, because many patrons decide not to wait in long lines for concessions. If they overestimate the attendance the theatre wastes money, because some employees sit around with nothing to do. Overall the theatre is profitable. Nevertheless, Ty and Chris have calculated that the theatre loses €1 for every unit of attendance by which the forecast misses the correct answer, whether it's an overestimate or an underestimate. On the following page are the forecasts that Ty and Chris made separately for the last 10 days. The correct attendance levels are also given. Study their forecasts carefully for a minute or two. Afterwards, we will ask you several questions about forecast accuracy.

Participants then saw a list of forecasts for Ty and Chris accompanied by the true attendance level. In the sequential format, these lists of estimates appeared on separate pages. In the simultaneous format, these lists of estimates were columns in the same table. In all cases, daily forecasts were generated for Ty and Chris assuming normally distributed random errors, where Ty had a better MAD than did Chris. Two conditions were used, one in which Ty was substantially more accurate than Chris (MADs of 31 and 47, respectively) and one in which Ty was somewhat more accurate than Chris (MADs of 34 and 42). In all cases, averaging produced a MAD lower than both individuals' MADs.

Three bracketing rate conditions were created by varying the patterns of error: No correlation in random error (40% rate), negative correlation in random error (80% rate), and opposing biases with no correlation in random error (80% rate). The actual stimuli for the three bracketing rates are shown in Figure 2 (small difference in MAD, simultaneous format). Averaging led to more improvement in conditions with more bracketing. The overall design crossed format (2) by difference in MAD (2) by bracketing rate (3). (The MAD manipulation produced no main effects or interactions for the main dependent variable, and will not be considered further.)

After studying the pattern of judgments, participants were told:

> In answering the questions below, please do not go back and re-examine the forecasts on the preceding pages. Rather, base your answers on the intuitive impressions you have already developed. In answering the questions, recall that the theatre loses €1 for every unit of attendance by which the forecast misses the correct answer, whether it's an overestimate or an underestimate.

Participants were then asked to estimate the MAD for the following three strategies (two additional strategies were included as filler):

Strategy 1. If Ty's estimate alone were always used, how much money would the theatre have lost per day on average? (*Ty alone*)

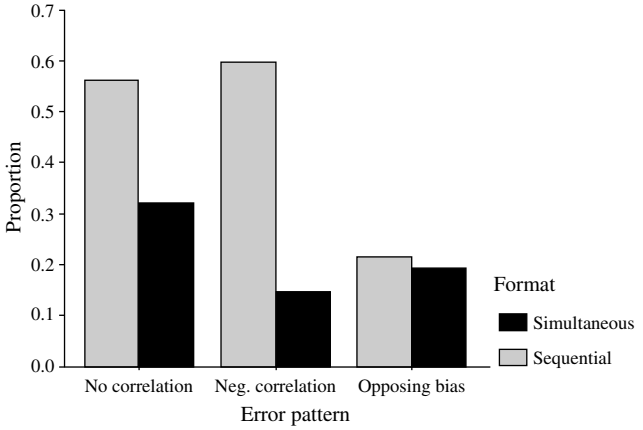
Strategy 2. If Chris's estimate alone were always used, how much money would the theatre have lost per day on average? (*Chris alone*)

Strategy 3. If the mean (that is, the midpoint) of Ty and Chris' estimates were always used, how much money would the theatre have lost per day on average? (*Averaging*)

Results and Discussion

For each participant, an *average judge score* (j) was calculated from their estimates for *Ty alone* and *Chris alone*. Participants' estimates for the averaging strategy were coded as *no better than the average judge* if their estimate was equal to or greater than j.

Figure 3 **Proportion of Participants in Experiment 2 That Expected Averaging to Be No More Accurate Than the Average Judge (Displayed by Error Pattern and Learning Environment)**



The rate at which participants mistakenly estimated that averaging would perform no better than the average judge varied systematically by condition (see Figure 3). As predicted, misappreciation of averaging was more likely in the sequential format than in the simultaneous format condition for the no correlation condition (proportions of 0.57 versus 0.33, $n = 87$, $p < 0.02$, Fisher's exact test) and for the negative correlation condition (0.60 versus 0.15, $n = 87$, $p < 0.001$, Fisher's exact test). Also, as expected, misappreciation of averaging was rare for the opposing biases condition, and did not differ by format (sequential 0.21 versus simultaneous 0.19, $n = 89$, ns). Also as before, participants were less likely to make the mistake in the negative correlation condition than in the no correlation condition, but only within the simultaneous format condition, where the degree of correlation was transparent (0.15 versus 0.33, $n = 90$, $p < 0.03$, Fisher's exact test).[3] As in Experiment 1, a large majority (90%) of those who believed that averaging would perform better than the average judge expected it to perform better than both judges. Also, as before, a large majority (84%) of those who believed that averaging would perform no better than the average judge expected it to perform exactly equal to the average judge.

In sum, participants reasoned much less accurately about averaging in sequential formats—in which concrete instances of bracketing were never directly observed—than in simultaneous formats. This environment effect, however, did not hold for opposing biases, which could be encoded at the individual level even in the sequential format. Experiment 2 confirmed that specific learning environments (simultaneous formats) and forms of bracketing (opposing biases) helped people to reason extensionally about averaging. In the General Discussion at the end of this paper, we consider whether these helpful circumstances are common.

[3] We also calculated the mean percentage improvement that participants expected from averaging (compared to *j*, the performance of the average judge). Expected improvement was large and sensitive to bracketing in the simultaneous format condition ($M_{\text{no correlation}} = 23\%$ versus $M_{\text{negative correlation}} = 43\%$); it was smaller and insensitive to bracketing in the sequential format condition ($M_{\text{no correlation}} = 19\%$ versus $M_{\text{negative correlation}} = 18\%$). Expected improvement from averaging was large for opposing biases in the simultaneous ($M = 45\%$) and sequential ($M = 50\%$) conditions.

Experiment 3

The first two experiments have shown that people commonly assume that averaging performs no better than the average judge and that certain factors, such as transparency and degree of bracketing, reduce this misperception. Both experiments, however, relied heavily on numerical information and required numerical responses. To the extent that calculation is difficult or distracting, it may lead to greater reliance on heuristic processing. In addition, numerical responses may invite a form of reasoning by representativeness (Tversky and Kahneman 1983) in which people equate the performance of averaging with the performance of the average judge. Although these mechanisms are real sources of misappreciating averaging, the effect may be of less interest if it depends exclusively on these mechanisms. Experiment 3 was designed to minimize numerical information to test people's *conceptual* understanding of averaging. To do so, the two "component" judges were described as being equally close to the truth in their forecasts. Participants were then asked to judge whether, over a series of judgments, "midpoint" estimates would be closer to the truth, further, or the same distance as both judges. Given some amount of bracketing, the midpoint strategy must be more accurate than both individual judges. This design reduced the use of numerical information in the stimuli and no longer asked participants to give a quantitative estimate of the effect of averaging.

Also, unlike the first two experiments, Experiment 3 elicited assumptions about bracketing rather than manipulating bracketing. Perhaps part of people's difficulty in recognizing the benefits of averaging is that they do not attend to external information about bracketing, but are nevertheless sensitive to their own assumptions. This allowed us to measure people's spontaneous assumptions about bracketing in the absence of external information and to test how they used it in their own reasoning. Also, by counterbalancing when the bracketing rate was elicited, we could test whether directing participants' attention to bracketing prior to evaluating averaging might enhance their understanding of its beneficial effects on accuracy.

As we found in Experiments 1 and 2, we expected that people would often incorrectly predict that averaging judgments would be no more accurate than the average judge. However, as in the previous experiments, we also expected that estimates would be more accurate as bracketing rate—in this case, imagined rate—increased.

Method

Participants. One hundred forty-nine participants (89% students, mean age = 22.4) were recruited at the Sorbonne in Paris to participate at the INSEAD Social Science Research Center. In exchange for participating in two short studies (about 15 minutes total), participants received a coupon for food at a nearby cafe worth €2.20. The present study was always completed first.

Materials and Procedure. Participants read the following scenario in French:

> Carl, Guy, and Pierre work at an art auction. They have a game that they play amongst themselves. Every time a painting is auctioned off, they each guess the selling price of the painting. Carl and Guy are equally good at guessing prices; sometimes Carl is closer and sometimes Guy is closer, but neither has a definitive advantage. Pierre always gets to guess last. He is trying out a new strategy, which is to always guess the midpoint between Carl's guess and Guy's guess. For example, if Carl guesses €1,000 and Guy guesses €1,400, Pierre would guess €1,200.

Participants' assumptions about bracketing were elicited with the following question:

> Suppose that the game is played for 100 paintings. For how many paintings would you expect the true selling price to be in between the guesses of Carl and Guy? (For example, suppose that the guesses of Carl and Guy are €300 and €500. If the selling price is greater than €300 and less than €500, then it is in between. Otherwise, it is not in between.)

Participants then estimated the number of 100 paintings for which the truth would fall *in between* and *not in between* Carl and Guy's guesses.

Participants' predictions about the effect of averaging were measured using a multiple choice format

> Compare Carl, Guy, and Pierre. Whose guesses will be closer on average to the true selling price (circle one)?
>
> (a) Carl and Guy will be equally close. Both will be closer than Pierre, on average.
>
> (b) Carl and Guy will be equally close. Pierre will be closer than both of them, on average.
>
> (c) Carl, Guy, and Pierre will be equally close, on average.
>
> (d) Carl and Pierre will be equally close. Both will be closer than Guy, on average.

The four options were varied in a Latin square design. There was no effect of order on response choice.

The order in which participants answered the bracketing rate question and strategy question was counterbalanced (bracketing first versus bracketing second). This allowed us to test whether thinking about bracketing prior to evaluating strategies improved understanding of averaging.

Results and Discussion

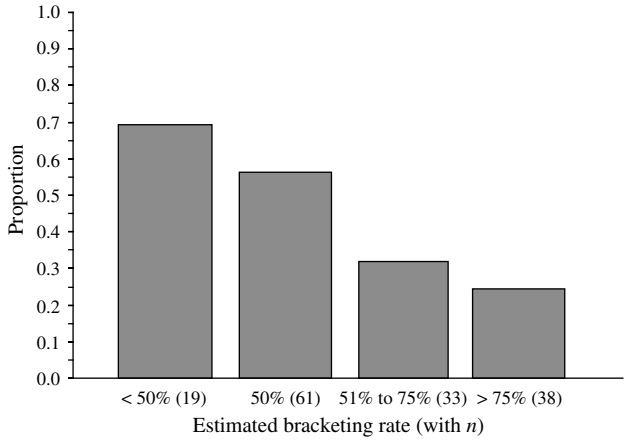
Participants estimated that the guesses of the component judges, Carl and Guy, would bracket the truth on 61% of trials, a rate that did not vary by question order ($M_{\text{bracketing first}} = 59\%$ versus $M_{\text{bracketing second}} = 62\%$, ns). Notably, all participants expected some bracketing.

Fifty percent of participants believed that averaging would be more accurate than both component judges (option (b)), 32% believed it would be equal to both (option (c)), and 13% believed it would be worse than both (option (a)). The remaining participants chose option (d) above, which does not have a clear interpretation. This pattern did not differ by question order (*bracketing first* = 47%, 34%, 16% versus *bracketing second* = 53%, 31%, 10%, respectively, ns). The modal participant correctly expected the midpoint strategy to perform better than the component judges. However, the basic proportion of participants who misunderstood the effect of averaging remained similar to those in Experiments 1 and 2: 45% of participants expected averaging to perform equal to (option (c)) or worse than (option (a)) the average judge.

Participants were sensitive to (or coherent with) their assumptions about bracketing rate. Participants who assumed a lower rate of bracketing were more likely to predict that averaging would perform no better than the average judge (Kendall's tau-b = -0.28, $p < 0.001$), and this held for both question orders (*bracketing first*—Kendall's tau-b = -0.22, $p < 0.02$, and *bracketing second*—Kendall's tau-b = -0.35, $p < 0.001$). Figure 4 displays this relationship by grouping participants into four roughly equal categories based on their bracketing rate assumptions (the modal 50% response could not be divided). As may be seen, a majority of participants expected bracketing to be 50% or less, of which a majority incorrectly expected averaging to perform no better than the average judge. However, a substantial minority expected bracketing rates to be higher than 50%, of which a large majority correctly predicted the effect of averaging.

Experiment 3 minimized numerical information and numerical responses to isolate people's conceptual understanding of averaging. The modal respondent correctly predicted the effect of averaging, although a similar proportion to Experiments 1 and 2

Figure 4 **Proportion of Participants in Experiment 3 That Expected Averaging to Be No More Accurate Than the Average Judge (Displayed by Participant's Estimate of the Bracketing Rate)**



incorrectly predicted that averaging would perform no better than the average judge. The change in method did not produce qualitatively different conclusions from the previous designs. Participants also reasoned in accord with their assumptions about bracketing rates: The higher their assumed bracketing rate, the more likely they were to predict that averaging would perform better than the average judge.

It is interesting to note two points about participants' bracketing rate assumptions. First, 82% of participants assumed bracketing rates of 50% or higher, and over 25% assumed bracketing rates of 75% or higher. Judges would tend to bracket the truth 50% of the time if they were unbiased and had uncorrelated random error, and would consistently bracket at a higher rate only if they had opposing biases, negatively correlated random errors, or both. In reality, because judges do share biases and have positively correlated random error in many judgment domains, 40% is at the high end of the range of empirically-observed bracketing rates. For example, Soll and Larrick (2005) observed mean bracketing rates between 20% and 43% across a variety of judgment tasks. A second point to note is that, because Experiment 3 measured but did not manipulate bracketing rates, it essentially identified "aggregation optimists" (who assumed both high bracketing and high benefits of averaging) and "aggregation pessimists" (who assumed ordinary bracketing and less benefit from averaging). This suggests the possibility that some people exhibit a compensatory bias: When bracketing is unobserved, a subset of people may assume an extreme but uncommon rate of bracketing. For such a group, receiving feedback on *representative* bracketing rates could have the ironic effect of discouraging their enthusiasm for averaging. We return to the reasoning of "aggregation optimists" in the General Discussion.

Overall, participants in Experiment 3 were somewhat more accurate about the consequences of averaging than were participants in the previous experiments. The improvement in performance may have been due to the change in population, in method, or both. In a follow-up study using a different sample from the same Sorbonne student population ($n = 64$), we presented participants with the Carl-Guy scenario followed by a 7-point response scale (*Pierre will be* $\ldots 1 =$ *much less accurate than Carl and Guy*, $4 =$ *equally accurate to Carl and Guy*, $7 =$ *much more accurate than Carl and Guy*) in place of the multiple-choice options in Experiment 3. On this scale, the mean response was 4.44, which was higher than the midpoint of 4 ($t(63) = 1.965$, $p = 0.055$). However, the median response was 4, with 14 participants predicting that averaging would be less accurate than the average judge, 21 predicting it would be the same, and 29 predicting it would be more accurate. Thus, as in Experiments 1 and 2, a narrow majority (55%) misunderstood the effects of averaging and a large minority (45%) had an accurate view.[4] All told, these results suggest that the student populations used in these studies did not reason reliably better (or worse) than the MBA samples used in Experiments 1 and 2.

We have observed in other research (Soll and Larrick 2005) that people frequently do not average when revising their own judgments, but choose to stay with their initial estimate, and we have proposed that misappreciation of averaging is one reason for the "choosing" strategy. To examine this point, the follow-up study included a second task in which participants estimated the number of marbles in a bottle, silently exchanged estimates with a partner, and then provided a final estimate that was rewarded for accuracy. Participants' beliefs about the benefits of averaging in the Carl and Guy scenario (measured on the 7-point scale) were positively related to the likelihood that they changed their initial marble estimates after seeing their partner's judgment (Kendall's tau-b $= 0.245$, $p < 0.03$). Of those participants who believed that averaging outperforms the average judge (gave a response of 5 or above in the Carl and Guy scenario), 59% revised their initial marble estimate toward their advisor, compared to 34% of the remaining participants.

Experiments 1 through 3 demonstrated that people often believe that averaging the estimates of two judges is no more accurate than the response of the average judge. They also demonstrated that the propensity to make this reasoning mistake was diminished when bracketing was common and made

[4] Mean estimated bracketing rate was 58.9%. Higher bracketing estimates were once again related to more accurate beliefs about the benefit of averaging (as measured on the full 7-point scale) (Kendall's tau-b $= 0.22$, $p < 0.03$).

transparent, thereby facilitating correct extensional reasoning. In the introduction, we suggested that failure to recognize the benefits of averaging were due to failures in extensional reasoning and to flawed inferential rules. Experiment 4 was designed to test the influence of these sources directly.

Experiment 4

Experiments 1 through 3 investigated an environmental factor—the transparency of bracketing—that made it is easier to reason correctly about the effect of averaging. What, however, underlies the basic failure to understand averaging? In the introduction, we proposed that people may have difficulty reasoning extensionally and may therefore substitute flawed inferential rules to predict the effect of averaging. Experiment 4 was designed to assess the joint contribution of extensional reasoning ability and flawed inferential rules measured at the level of the individual judge.

The benefits of averaging may be appreciated at a deep level by careful extensional reasoning—by imagining the space of possible outcomes and their implications. There is evidence, however, that not all people are equal in their ability to reason extensionally (Stanovich and West 2000). In an influential collection of studies, Stanovich and West found reliable individual differences in people's ability to make normative decisions across a wide range of problem domains. Following Stanovich and West, we measured individual differences in extensional reasoning ability by testing performance on two reasoning problems that are unrelated to our domain. We expected to find that performance on these unrelated problems would predict success at understanding the effect of averaging.

We have argued that in place of extensional reasoning, people may rely instead on flawed inferential rules and aphorisms that are acquired from the larger culture (Fiske et al. 1998, Nisbett and Ross 1980) or induced through experience (Hogarth 2001, Holland et al. 1986). For example, research on cross-cultural differences has found that members of individualistic countries are more likely than those in collectivistic countries to endorse the statement, "Decisions made by individuals are usually of higher quality than decisions made by groups" (Hofstede 1984, p. 160). Similar differences in individualism and collectivism occur at the individual level within countries (Oyserman et al. 2002, Singelis 1994, Triandis 1995). We suspect that individualists understand group decisions differently than do collectivists. When different individuals hold conflicting opinions in a group, individualists attempt to identify the most able person and choose his or her judgments, which we have termed "chasing the expert" (Soll and Larrick 2005, see also Surowiecki 2004). The individualist's fear is that pristine individual judgment will be influenced and distorted by the group's response; conforming to others by compromising or "splitting the difference" is seen as abdicating the truth (Nisbett et al. 2001).[5] As one popular group decision-making textbook (Fisher and Ellis 1990 p. 276) instructs, "Compromise should be considered a last-resort measure . . . . [E]verybody loses something." This is sound advice in some contexts, such as multiple-issue negotiations (Raiffa 1982), but it does not generalize to combining estimates. By contrast, collectivists may be more willing to value compromise and to yield to it (Nisbett et al. 2001).

Although culture may provide basic beliefs, such as "compromise leads to mediocrity," there are often aphorisms suggesting the opposite conclusion, such as "the truth lies in the middle." Which rules will people hold and use? Culture is one determinant. Another determinant may be a person's statistical sophistication. Finally, experience in the right environment may provide the opportunity to induce correct rules (Hogarth 2001, Holland et al. 1986). Regardless of why people favor one aphorism over another, we expect that people who endorse aphorisms that support compromise are more likely to predict that averaging will be beneficial (which is distinct from understanding why it is beneficial).

Overview of Specific Measures. We measured extensional reasoning ability by testing performance on two reasoning problems that are formally unrelated to combining opinions: The Will Rogers effect (Messick and Asuncion 1993) and Simpson's paradox (Simpson 1951). The Will Rogers effect (WRE) is named after the famous humorist who once observed that the migration of Oklahomans to California during the Great Depression raised the average intelligence of both states. Our WRE item (Question 2 in the appendix) asked about the possible effects of moving a group of employees from one department in a company to another department. One combination (2b) stated, "The percentage of men in Department A increases and, at the same time, the percentage of men in Department B increases." Although the correct extensional analysis shows that this combination is logically possible, many people consider it impossible. They incorrectly expect that changes in the partitions

[5] Suspicion of group influence on individuals is clearly illustrated in the social psychology studies of the 1950s and 1960s, which focused on how conformity to group norms leads to "the loss of individuality, restriction of creativity, and reduction of all group members to the level of mediocrity" (Shaw 1976). Vernon Allen (1965, pp. 135–136) summed up the assumptions of this research by observing that "conformity has captured the interest of social psychologists and laymen alike and has been thoroughly censured by both."

(departments) must move in opposite directions to match the unchanging proportion in the population (company).

Simpson's paradox (SP) occurs when a relationship between two variables reverses once data are conditioned on a third (confounding) variable. Our SP question (Question 3 in the appendix) asked about the average salary of women when men had higher salaries within the two departments making up a company. One combination (3b) stated, "In the company as a whole, the average salary of the women is higher than the average salary of the men." Although the correct extensional analysis shows that this combination is logically possible, many people consider it impossible (for other examples, see Curley and Browne 2001 and Fiedler 2000). Once again, they incorrectly expect that properties of partitions (departments) must match properties of the population (company).

To assess inferential rules, we asked participants to evaluate two cultural aphorisms that favor averaging (e.g., "the truth lies in the middle") and two that did not (e.g., "compromise leads to mediocrity"). To measure reasoning about averaging, we asked participants to judge different hypothetical outcomes of averaging (see Questions 1 and 4 in the appendix). We expected that both the ability to reason extensionally and endorsement of pro-averaging aphorisms would be positively related to correct inferences about averaging. We also wanted to test whether the two predictors interacted—did they compensate for each other? Were they jointly necessary? Although cultural aphorisms tend to point to a general conclusion about averaging as good or bad, they are otherwise imprecise. We expected that reasoning by aphorisms would not be as accurate as applying a full extensional analysis.

Method

Participants. Participants were 60 students at Duke University who were paid $6 for taking part in a 20-minute study.

Materials and Procedure. Participants read a scenario about a student group who ran a university film series. They were told that one of the group's tasks was to meet and to forecast attendance for different films they were considering showing. Participants read:

> The actual attendance at the film series varies between 50 and 250 depending on the popularity of the specific movie. Historically, the five members of the group have varied in how accurate their predictions are. No member of the group makes perfect predictions—that is, each member sometimes guesses too high and sometimes too low. However, some members are more accurate than others. The group measures accuracy in terms of absolute difference from the truth; they only care about deviation from the truth, not whether it is high or low. Thus, if the true attendance is 120, then someone who guesses 110 misses by 10, someone who guesses 130 misses by 10, and they are equally inaccurate. Over the last 50 movies, the most accurate member of the group, Tracy, has misestimated the actual attendance by 20 people on average (sometimes high, sometimes low). By comparison, the least accurate member, Kelly, has misestimated the actual attendance by 30 people on average (sometimes high, sometimes low). The other members of the committee are less accurate than Tracy but more accurate than Kelly.

Participants were then asked to think about how the different forecasts from each member of the group might be used. This was followed by the question, "Think about the 'Forecasting Committee' described above. In your opinion, how true is each of the following sayings in this situation?" Each saying was rated on a separate 1 to 7 scale, with the endpoints *not at all true* and *very true*. The sayings were (a) "Two heads are better than one," (b) "True expertise is knowing when to defer to an expert," (c) "The truth lies in the middle," and (d) "Compromise leads to mediocrity."

Next, participants answered the questions that appear in the appendix. Two questions were designed to measure extensional reasoning about the effect of averaging on absolute deviation (Question 1 regarding a single judgment and Question 4 regarding multiple judgments) and two were designed to measure extensional reasoning in other domains (Question 2 on WRE and Question 3 on SP). The specific instructions for these questions were, "For each of the scenarios below, tell us whether you think that the scenario is **possible** (i.e., it could be true under some circumstances) or **impossible** (i.e., it could never be true). Please try to be as accurate as possible in your answers." The order of questions was counterbalanced and had no effect on the results.

Results and Discussion

The aphorism questions correlated reasonably well with each other (average $r = 0.33$ after reverse coding items (b) and (d)) and were averaged to create an aphorisms score for each participant. High scores on the aphorisms measure reflected a general endorsement of the benefits of aggregation. The average participant had a score near the midpoint on this scale ($M = 4.30$, $sd = 1.19$). The number of correct responses to Question 2 (WRE, four parts) and Question 3 (SP, two parts) were converted to a proportion correct to create an extensional reasoning measure ($M = 0.68$, $sd = 0.16$). The aphorisms measure was positively correlated with the extensional reasoning measure ($r = 0.26$, $p < 0.05$), suggesting that general faith in averaging could be partly a product

Table 3 Regression Predicting Reasoning About Averaging Opinions (Proportion Correct) from the Aphorisms Measure and the Extensional Reasoning Measure (Experiment 4)

Predictor	Equation 1	Equation 2	Equation 3	Equation 4
Constant	0.80*	0.80*	0.80*	0.79*
Aphorisms measure	0.05*		0.04	0.04
Extensional reasoning measure		0.04	0.03	0.02
Interaction				0.05*
Adjusted R^2	0.05*	0.03	0.05	0.11*

Note. The aphorisms and extensional reasoning measures were standardized. The interaction term is the product of the aphorisms measure and the extensional reasoning measure. The coefficients describe how a one standard deviation change in a predictor variable affects proportion correct. All standard errors were 0.025. *$p < 0.05$, two-tailed.

of extensional reasoning ability (with the caveat of interpreting a correlation causally). Finally, the number of correct responses to Questions 1 and 4 (three parts each) were converted to a proportion correct for each participant to create an averaging opinions measure ($M = 0.80$, $sd = 0.20$). This measure served as the dependent variable. The averaging opinions measure was significantly correlated with the aphorisms measure ($r = 0.25$, $p = 0.05$) but not with the extensional reasoning measure ($r = 0.21$, $p = 0.11$) (see also Equations (1) and (2) in Table 3).

To test how the aphorisms measure and extensional reasoning measure jointly predicted the averaging opinions measure, we regressed the averaging opinions measure on both predictors and an interaction term. To minimize collinearity between the predictors and the interaction term, we mean-centered the measures by converting them to z-scores before taking their product. As desired, the resulting interaction term was not significantly correlated with either predictor (with aphorisms, $r = 0.07$, ns; with extensional reasoning, $r = 0.16$, ns). Four regression equations are reported in Table 3.[6] The full model (Equation (4)) revealed a significant interaction. To illustrate this interaction, we performed a median split on the aphorisms and extensional reasoning measures, and found that participants who were above the median on both predictors had a mean score of 0.91 on the averaging opinions measure. The remaining participants performed worse. Those who were below the median on both had a mean score of 0.76 and those who were above on one but not on the other had a mean score of 0.74 (high extensional, low aphorism) and 0.77 (high aphorism, low extensional). These data suggest that, to reason accurately about the averaging principle, people need both faith in the benefits of aggregation *and* the ability to think extensionally about possible outcomes.

We performed additional analyses to shed light on what people are thinking when they misappreciate the effects of averaging. First, the averaging opinions questions (Questions 1 and 4) can reveal at a more fine-grained level the specific mistakes people make. Consider Question 4, where participants universally answered part (b) correctly and 28 of 60 participants answered all parts correctly. The remaining erroneous responses to parts (a) and (c) can be used to distinguish between different types of reasoning mistakes. Eight people responded that both (a) and (c) were impossible, implying that averaging can lead to *only* average performance. Six people responded that (a) was impossible—averaging can perform *no better* than the average; and 18 responded that nothing was impossible—averaging can perform *better or worse* than average. As these results show, the majority of those who misappreciated averaging in Experiment 4 believed that everything was possible—good, bad, and no change. This suggests that, in Experiments 1 through 3, participants who equated averaging with the performance of the average judge were often reporting an expected value in a subjective distribution of outcomes.

A final analysis of the averaging opinions questions (Questions 1 and 4) revealed that people performed similarly across the two questions. The accuracy score on one question was highly correlated with the score on the other ($r = 0.63$, $p < 0.001$). This pattern suggests reliable individual differences in the tendency to appreciate (or misappreciate) averaging.

General Discussion

Under a range of conditions, a majority of people erroneously believed that averaging the estimates of two judges would perform no better than the average judge. Performance at the level of the average judge is the worst possible outcome for averaging, and occurs only when judges *never* bracket the truth. Because the judges in our stimuli bracketed the truth between 24% and 90% of the time, averaging was always superior to the average judge (and, in these cases, superior to both judges). Across a range of methods and tasks, a subset of people—in many cases a majority—did not reason with this most basic principle of aggregation.

Although participants frequently mispredicted that averaging would perform no better than the average judge, factors that facilitated extensional reasoning

[6] The resulting coefficients show how a standard deviation change in a predictor affects proportion correct. Given that the dependent variable had only five levels, the ordinary least squares (OLS) regression was repeated as a multinomial logistic regression. This analysis yielded stronger significance levels: The fourth model was significant at $p < 0.01$ and the interaction at $p < 0.005$. The main effects remained nonsignificant. Because the qualitative results were unaffected, we report the OLS results for ease of interpretation.

greatly mitigated this mistake: Participants reasoned more accurately when the bracketing rate was high and when bracketing was made transparent (either through observing judges simultaneously or through the ease of recognizing opposing biases).

Unfortunately, there are reasons to believe that many of the facilitating conditions we identified might be uncommon in daily life. First, in many cases only summary statistics on accuracy are available. For example, individual investors rarely track the performance of multiple stock analysts on a stock-by-stock basis. Rather, investment services provide records for multiple analysts, showing how well investors would have performed had they followed the advice of each. Information that might help people see the benefit of averaging, such as interanalyst error correlations or bracketing rates, are typically not provided. Second, even when estimates of multiple judges are presented simultaneously, they may not be presented with the truth. For example, in the days leading up to a football Sunday, many newspapers publish sportswriters' predictions of victory margins (e.g., Packers by 4). However, after the games have been played, they do not publish charts showing all forecasters' predictions together with the game outcomes. If they did, instances of bracketing might "leap out" of the data as in Experiment 2, making the power of averaging more salient. Instead, newspapers often summarize the accuracy of each sportswriter at season's end. Life abounds with data on *individual* accuracy and error but rarely reveals *interpersonal* patterns of error that make bracketing transparent.

One circumstance in which participants did recognize the benefits of averaging was when judges exhibited opposing biases, which may be a fairly common event. For example, a dieter may learn that the bathroom scale tends to overestimate weight and the bedroom scale tends to underestimate. We might expect the dieter to start averaging the two values. However, biases may not always be obvious. Ten Wall Street investment firms, including Citigroup and Merrill Lynch, recently agreed to pay a combined penalty of nearly $1 billion to settle a lawsuit charging that stock analysts' forecasts have been overly optimistic for many years (White 2002). Apparently a small systematic bias can easily persist undetected by average investors, even in a data-rich, heavily scrutinized environment.

How might reasoning about averaging be improved? We believe that with repeated exposure to instances of bracketing, judges can induce the general rule that averaging will perform better than the average judge. (Clear evidence that judges have generalized the rule would include their applying it to new problems when bracketing is unobservable.) Substantially more training would be needed, however, to help people gain a deep understanding of the parameters that determine the *degree* of benefit from averaging. Averaging will not always surpass the performance of both judges, such as when bracketing rates are low or when the judges differ greatly in accuracy (Soll and Larrick 2005). Training people to understand how correlated random error, shared bias, and differences in accuracy jointly determine the performance of averaging relative to the better judge is a more daunting task. This inference, however, is the critical one that determines whether one should average judgments in a pair or choose the more expert judge (Soll and Larrick 2005).

The distinction between a general belief in the benefits of averaging and a deep, extensional understanding of averaging is useful for analyzing the steady minority of our participants who believed that averaging would outperform *both* judges. This group is not likely to be homogeneous. Some are likely to be sophisticated extensional reasoners about averaging. Others, however, may be "aggregation optimists" who believe in the benefits of averaging as an article of faith.[7] This second group may have too much faith if they assume that bracketing rates are routinely around 75% (as seen in Experiment 3) or that averaging always performs better than both judges in a pair. Consequently, they may elect to average judgments in situations when choosing the estimates of a single expert would be more accurate.

Future research should explore the connection between beliefs about averaging and the strategies judges apply when using advice to revise their estimates. As demonstrated in the follow-up study to Experiment 3, participants who are more optimistic about averaging are more likely to revise their estimates in the direction of advice. Their willingness to use advice may reflect an understanding of averaging, or it may reflect other factors, such as risk aversion (Soll and Larrick 2005) or the availability of new reasons (Yaniv and Kleinberger 2000).

We opened this paper by describing an old debate on why groups produced more accurate judgments than individuals, in which intangible group qualities (the "Creative Plus") were contrasted with statistical explanations. We close with a similar, contemporary issue. Quantitative group exercises are routinely run in MBA classes in which the accuracy of final group judgments, arrived at through discussion and debate, is contrasted with the accuracy of the average individual's initial judgments. The improvement in

[7] We do not mean to imply that optimism is a personality difference, but simply a difference in assumptions that individuals make on a given aggregation task. An interesting question is whether there are stable differences in the assumptions that people hold and evoke for aggregation tasks.

performance is frequently offered as clear evidence of the power of discussion and debate in groups. However, if the group is doing anything resembling averaging and there is any bracketing, it is a statistical necessity that final group judgments will be superior to average individual judgments. Without belaboring the point, the average individual is an uninformative benchmark for this type of task. The more informative benchmark is whether the group's answers can outperform the strategy of simply averaging initial individual estimates. Empirically, interacting groups tend to perform at the level of mechanical averaging (Hastie 1986), presumably because normative and informational influence (Deutsch and Gerard 1955) lead to compromise.

The fact that groups perform at the level of averaging suggests the radical prescription that groups might wish to forego discussion and simply average their private quantitative judgments, both for efficiency reasons and to preserve independence in judgment. When we propose this strategy to our students, however, they resist it (cf. Kleinmuntz 1990). Partly they argue that there are benefits from group discussion that are not realized from mechanical combination, including building cohesion, commitment to the decision, and institutional memory. But they also argue that averaging leads to mediocrity and that they need discussion to identify who is the true expert. We have had hundreds of students take part in quantitative group judgment tasks, and the overwhelming majority has (privately) predicted that their group answers would be more accurate than averaging initial individual judgments. True to the results in the literature (Hastie 1986), however, half the groups fall short of this benchmark. The preference for discussion over averaging is not in itself harmful, because there are social benefits that offset the time costs. The real danger is that groups reduce the accuracy of their judgments in this type of task when they "chase the expert" and fail to appreciate—and capture—the benefits of compromise.

Conclusion

Averaging is a powerful way to reduce error across many settings, including combining the opinions of experts (Clemen 1989, Hogarth 1978), integrating the judgments of group members (Einhorn et al. 1977, Gigone and Hastie 1997), and revising one's own opinion (Soll and Larrick 2005). Yet people often do not take advantage of the benefits of averaging. We believe that these studies identify one of the major reasons people fail to exploit averaging—many hold an incorrect theory about the effect of averaging, believing that it equals average performance. This erroneous belief, in combination with overconfidence in the ability to identify more expert judges, leads people to focus on finding the perceived expert and to rely on that expert's judgment. The failure to combine judgments comes at a high price in many common social and organizational settings.

Acknowledgments

The authors thank David Budescu, Bob Clemen, Tom Wallsten, Bob Winkler, Ilan Yaniv, and one anonymous reviewer for their feedback on this paper, and Enrico Diecidue, Spyros Makridakis, Ayse Onculer, and Valeria Noguti for providing opportunities for data collection. Financial support provided by the INSEAD research and development department is gratefully acknowledged. They thank participants at the Behavioral Decision Research in Management Conference in Chicago (June 2002), at the IFORS Conference in Edinburgh (July 2002), the Tilburg Summer Symposium on Economics and Psychology (September 2002), and the Conference on Information Aggregation at the University of Maryland (May 2003) for their many helpful comments.

Appendix

(1) (Absolute Deviation Question). John, Isabelle, and Karla predicted the year end price of stock ABC. John missed the true price by \$10 and Isabelle missed by \$5. Karla overheard the forecasts of John and Isabelle and predicted the average of their forecasts. (Please circle a response for (a), (b), and (c).)

(a) Karla missed the true price by less than \$5. **Possible Impossible**

(b) Karla missed the true price by \$7.50. **Possible Impossible**

(c) Karla missed the true price by \$10.00 or more. **Possible Impossible**

(2) (Will Rogers Effect Question). A company has two departments, Department A and Department B. Department A has 60% men and Department B has 40% men. A group of employees is transferred from Department A to Department B. No other employees are moved. (Please circle a response for (a), (b), (c), and (d).)

(a) The percentage of men in Department A increases and, at the same time, the percentage of men in Department B decreases. **Possible Impossible**

(b) The percentage of men in Department A increases and, at the same time, the percentage of men in Department B increases. **Possible Impossible**

(c) The percentage of men in Department A decreases and, at the same time, the percentage of men in Department B increases. **Possible Impossible**

(d) The percentage of men in Department A decreases and, at the same time, the percentage of men in Department B decreases. **Possible Impossible**

(3) (Simpson's Paradox). All employees in a company work in one of two departments, Department A or Department B. Each department contains some men and some women. In Department A, the average salary of the men is higher than the average salary of the women. In Department B, the average salary of the men is higher than the

average salary of the women. (Please circle a response for (a) and for (b).)

(a) In the company as a whole, the average salary of the men is higher than the average salary of the women. **Possible Impossible**

(b) In the company as a whole, the average salary of the women is higher than the average salary of the men. **Possible Impossible**

(4) (Mean Absolute Deviation Question). Jill, Steve, and Bill predicted the total number of points scored by their favorite basketball player, Alana Beard, in 20 different basketball games. They each tried to be as close as possible to the truth. They measured "closeness" by taking the difference between their guess and the truth. For example, if Beard scored 20 total points in a game, then a guess of 15 points would miss the truth by 5 points; similarly, a guess of 25 points would miss the truth by 5 points. Over the 20 games, Jill missed the true point total by 4 points on average (sometimes missing too high, sometimes missing too low); Steve missed the true point total by 6 points on average (sometimes missing too high, sometimes missing too low). Bill overheard the forecasts of Jill and Steve and predicted the midpoint of their estimates for each game. (Please circle a response for (a), (b), and (c).)

(a) Over the 20 games, Bill missed the true point total by 2 points or less on average. **Possible Impossible**

(b) Over the 20 games, Bill missed the true point total by 5 points on average. **Possible Impossible**

(c) Over the 20 games, Bill missed the true point total by 6 points or more on average. **Possible Impossible**

Correct answers: 1a P; 1b P; 1c I; 2a P; 2b P; 2c P; 2d I; 3a P; 3b P; 4a P; 4b P; 4c I.

References

Allen, V. L. 1965. Situational factors in conformity. L. Berkowitz, ed. *Advances in Experimental Social Psychology*, Vol. 2. Academic Press, New York, 133–175.

Ariely, D., W. T. Au, R. H. Bender, D. V. Budescu, C. Dietz, H. Gu, T. S. Wallsten, G. Zauberman. 2000. The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych.: Appl.* **6** 130–147.

Armstrong, J. S. 2001. Combining forecasts. J. S. Armstrong, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer, New York.

Brase, G. L. 2002. Ecological and evolutionary validity: Comments on Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Carverni's (1999) Mental Model Theory of extensional reasoning. *Psych. Rev.* **109** 722–228.

Bruce, R. S. 1935. Group judgments in the fields of lifted weights and visual discrimination. *J. Psych.* **1** 117–121.

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* **5** 559–609.

Clemen, R. T., R. L. Winkler. 1986. Combining economic forecasts. *J. Bus. Econom. Statist.* **4** 39–46.

Curley, S. P., G. J. Browne. 2001. Normative and descriptive analysis of Simpson's Paradox in decision making. *Organ. Behavior Human Decision Processes* **84** 308–333.

Dawes, R. M., B. Corrigan. 1974. Linear models in decision making. *Psych. Bull.* **81** 95–106.

Deutsch, M., H. B. Gerard. 1955. A study of normative and informational social influences upon individual judgment. *J. Abnormal Social Psych.* **51** 629–636.

Einhorn, H. J., R. M. Hogarth, E. Klempner. 1977. Quality of group judgment. *Psych. Bull.* **84** 158–172.

Eysenck, H. J. 1939. The validity of judgments as a function of number of judges. *J. Experiment. Psych.* **25** 650–654.

Fiedler, K. 2000. Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psych. Rev.* **107** 659–676.

Fildes, R., S. Makridakis. 1995. The impact of empirical accuracy studies on time series analysis and forecasting. *Internat. Statist. Rev.* **63** 289–308.

Fisher, B. A., D. G. Ellis. 1990. *Small Group Decision Making: Communication and the Group Process*, 3rd ed. McGraw Hill, New York.

Fiske, A. P., S. Kitayama, H. R. Markus, R. E. Nisbett. 1998. The cultural matrix of social psychology. D. T. Gilbert, S. T. Fiske, G. Lindzey, eds. *The Handbook of Social Psychology*, Vol. 2. McGraw Hill, New York, 915–981.

Gigone, D., R. Hastie. 1997. Proper analysis of the accuracy of group judgments. *Psych. Bull.* **121** 149–167.

Gonzalez, R. 1994. When words speak louder than actions: Another's evaluations can appear more diagnostic than their decisions. *Organ. Behavior Human Decision Processes* **58** 214–245.

Gordon, K. 1924. Group judgments in the field of lifted weights. *J. Experiment. Psych.* **3** 398–400.

Gordon, K. 1935. Further observations on group judgments of lifted weights. *J. Psych.* **1** 105–115.

Harvey, N., I. Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organ. Behavior Human Decision Processes* **70** 117–133.

Hastie, R. 1986. Experimental evidence on group accuracy. B. Grofman, G. Owen, eds. *Information Pooling and Group Decision Making*. JAI Press, Greenwich, CT, 129–157.

Hastie, R., T. Kameda. 2005. The robust beauty of the majority rule. *Psych. Rev.* **112** 494–508.

Hoffrage, U., S. Lindsey, R. Hertwig, G. Gigerenzer. 2000. Communicating statistical information. *Science* **290** 2261–2262.

Hofstede, G. 1984. *Culture's Consequences*. Sage, Newbury Park, CA.

Hogarth, R. M. 1978. A note on aggregating opinions. *Organ. Behavior Human Decision Processes* **21** 40–46.

Hogarth, R. M. 2001. *Educating Intuition*. The University of Chicago Press, Chicago, IL.

Holland, J. H., K. J. Holyoak, R. E. Nisbett, P. R. Thagard. 1986. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press, Cambridge, MA.

Johnson, T. R., D. V. Budescu, T. S. Wallsten. 2001. Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic values. *J. Behavioral Decision Making* **14** 123–140.

Kahneman, D., S. Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. T. Gilovich, D. Griffin, D. Kahneman, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press, New York, 49–81.

Kelley, T. L. 1925. The applicability of the Spearman-Brown formula for the measurement of reliability. *J. Ed. Psych.* **16** 300–303.

Kleinmuntz, B. 1990. Why we still use our heads instead of formulas: Toward an integrative approach. *Psych. Bull.* **107** 296–310.

Knight, H. C. 1921. *A Comparison of the Reliability of Group and Individual Judgments*. Master's thesis, Columbia University, New York.

Lewis, C., G. Keren. 1999. On the difficulties underlying Bayesian reasoning: A comment on Gigerenzer and Hoffrage. *Psych. Rev.* **106** 411–416.

Lorge, I., D. Fox, J. Davitz, M. Brenner. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psych. Bull.* **55** 337–372.

Messick, D. M., A. G. Asuncion. 1993. The Will Rogers illusion in judgments about social groups. *Psych. Sci.* **4** 46–48.

Nisbett, R. E., L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice Hall, Englewood Cliffs, NJ.

Nisbett, R. E., D. H. Krantz, C. Jepson, Z. Kunda. 1983. The use of statistical heuristics in everyday inductive reasoning. *Psych. Rev.* **90** 339–363.

Nisbett, R. E., K. Peng, I. Choi, A. Norenzayan. 2001. Culture and systems of thought: Holistic versus analytic cognition. *Psych. Rev.* **108** 291–310.

Oyserman, D., H. M. Coon, M. Kemmelmeier. 2002. Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psych. Bull.* **128** 3–72.

Preston, M. G. 1938. Note on the reliability and validity of the group judgment. *J. Experiment. Psych.* **22** 462–471.

Raiffa, H. 1982. *The Art and Science of Negotiation*. Harvard University Press, Cambridge, MA.

Ross, L., R. E. Nisbett. 1991. *The Person and the Situation: Perspectives of Social Psychology*. McGraw Hill, New York.

Shaw, M. E. 1976. *Group Dynamics: The Psychology of Small Group Behavior*, 2nd edition. McGraw-Hill, New York.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Statist. Social* (*Ser. B*) **13** 238–241.

Singelis, T. M. 1994. The measurement of independent and interdependent self-construals: An individual level analysis. *Personality Social Psych. Bull.* **20** 580–591.

Smith, M. 1931. Group judgments in the field of personality traits. *J. Experiment. Psych.* **14** 562–565.

Soll, J. B. 1999. Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psych.* **38** 317–346.

Soll, J. B., R. P. Larrick. 2005. Strategies for revising judgment: How, and how well, do people use others' opinions? Duke University, Durham, NC.

Sorkin, R. D., R. West, D. E. Robinson. 1998. Group performance depends on the majority rule. *Psych. Sci.* **9** 456–463.

Stanovich, K. E., R. F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral Brain Sci.* **23** 645–665.

Steiner, I. D. 1972. *Group Processes and Productivity*. Academic Press, New York.

Stroop, J. R. 1932. Is the judgment of the group better than that of the average member of the group? *J. Experiment. Psych.* **15** 550–562.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday, New York.

Triandis, H. 1995. *Individualism and Collectivism*. Westview, Boulder, CO.

Tversky, A., D. Kahneman. 1983. Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psych. Rev.* **90** 293–315.

von Winterfeldt, D., W. Edwards. 1986. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, UK.

Wallsten, T. S., D. V. Budescu, I. Erev, A. Diederich. 1997. Evaluating and combining subjective probability estimates. *J. Behavioral Decision Making* **10** 243–268.

Watson, G. B. 1928. Do groups think more efficiently than individuals? *J. Abnormal Social Psych.* **23** 328–336.

White, B. 2002. Wall Street agrees to mend its ways. *The Washington Post* (December 21) A1.

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

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