Intuitions about Combining Opinions: Misappreciation of the Averaging Principle

by

R. Larrick
and
J. Soll

2003/09/TM

Running Head: MISAPPRECIATION OF AVERAGING

Intuitions about Combining Opinions:

Misappreciation of the Averaging Principle

Richard P. Larrick

Duke University

Jack B. Soll

Associate Professor of Decision Sciences, INSEAD.

Word Count: 3,985 (including author notes)

Richard Larrick
Fuqua School of Business
Duke University
Box 90120
Durham, NC 27708

larrick@duke.edu
919 660-4076

Abstract

Averaging estimates is an effective way of improving accuracy when combining expert judgments, integrating group members' judgments, or using advice to modify personal judgments. If the estimates of two imperfect judges ever fall on either side of the truth, which we term bracketing, averaging must outperform the average judge. We hypothesized that people often hold an incorrect theory about averaging, falsely believing that the average of two judges' estimates would be no more accurate than the average judge. Experiment 1 confirmed that this misconception was common across a range of bracketing rates. Experiment 2 demonstrated that the effectiveness of averaging can be recognized when bracketing was made transparent. We conclude by describing how every day life provides few opportunities to learn the benefits of averaging and how misappreciating averaging contributes to poor intuitive strategies for combining estimates.

Intuitions about Combining Opinions:
Misappreciation of the Averaging Principle

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). 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). The ultimate conclusion, however, was that statisticized groups revealed nothing special about groups of people, but confirmed instead the statistical principle that aggregation of noisy 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, Fox, Davitz, & Brenner, 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. As a consequence, they often reason poorly about how best to use information from others to form or revise judgments.

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 judgments* (this is a special case of Jensen’s inequality; cf. Hogarth, 1978). To illustrate, imagine two people forecasting the high temperature in Honolulu tomorrow, which turns out to be 73º. If they guess 60º and 70º, they miss by 13º and 3º degrees respectively, or 8º on average. The average guess, 65º, also misses by 8º degrees. Thus, the average estimate performs no worse than the average judge. Now imagine they guess 60º and 80º, in which case the two estimates "bracket" the truth. In this instance, their guesses miss by 13º and 7º, or 10º on average. But the average guess of 70º misses by only 3º. Averaging outperforms the average individual (and, in this case, happens to outperform both individuals). When accuracy is measured by absolute deviation from the truth (or any weakly convex function), the averaging principle always holds. Averaging is even more advantageous if mean squared error is used, because in that case averaging outperforms the average individual even if both estimates err in the same direction.

An important implication of the averaging principle is that, over multiple judgments, averaging will outperform the average judge given at least one instance of bracketing. The actual rate of bracketing is likely to be much higher, and as it increases, so does the power of averaging. Two judges who have normally distributed errors that are mean-zero and uncorrelated will bracket the truth 50% of the time. If the judges are approximately similar in skill, averaging will improve accuracy by about 29%. The bracketing rate will be lower than 50% if judges share a bias (e.g., always guessing above the truth, such as 105º and 110º in the Honolulu example), or if they have positively correlated errors (e.g., underestimating in winter and overestimating in summer). Analogously, opposing biases and negatively correlated error will result in bracketing rates greater than 50%.

Extensive research in the forecasting (Armstrong, 2001; Clemen, 1989), decision making (Wallsten, Budescu, Erev, & Diederich, 1997), and groups (Einhorn, Hogarth, & Klempner, 1977; Gigone & Hastie, 1997) literatures has confirmed that averaging is a powerful and robust way of reducing error in judgment. 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 & Winkler, 1986; Fildes & Makridakis, 1995). In an extensive review, Armstrong (2001) reanalyzed thirty 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%.

Averaging itself is not critical to improving judgments. Weights in rough proximity to whatever weights are analytically optimal yield similar levels of improvement (von Winterfeldt & Edwards, 1973, as cited in Dawes, 1979). For example, if the appropriate weighting scheme is 50/50, using a 70/30 split will be nearly as accurate. However, our own empirical studies indicate that laypeople frequently do not combine at all, but instead choose between estimates (Soll & Larrick, 2003). In one study, participants first estimated salaries for graduates of 25 business programs. They were paid based on their revised estimates after viewing the responses of another participant. Although participants reduced error by 10% with their intuitive revision strategies, they would have improved by 16% had they consistently averaged. Because participants used extreme weighting schemes of 100/0 or 0/100 about half the time, they often missed out on the benefits of combination.

It could be that people avoid averaging because they are confident in their ability to identify the better judge. Such confidence may be misplaced but it upholds the romantic notion that one should strive for the very best, and not "settle" for averaging (cf. Kleinmuntz, 1990). But a second reason people might not average is that they simply do not recognize its effect. Specifically, we propose that many people mistakenly believe that averaging leads to average performance. Partly this belief reflects the application of an incorrect mathematical rule. However, individualistic norms in the West that equate compromise with mediocrity may also be a contributing factor.

This paper presents two experiments that examine people's beliefs about the effect of averaging judgments on reducing error. The first experiment tested people's understanding of averaging as an abstract principle. The second experiment investigated whether people could induce the benefits of averaging by observing sets of data. Of critical interest, we manipulated the bracketing rate, or 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. In general, we expected that the averaging principle would be difficult to grasp (the dozen empirical papers on averaging between the 1920s and 1940s are a testament to this difficulty!), but that the benefits of averaging would be more easily recognized when the bracketing rate is both high and salient.

Experiment 1

Assuming at least some bracketing, do people hold the abstract principle that averaging must outperform the average judge? We manipulated the bracketing rate for two hypothetical judges by varying correlation in error across three levels: Positive correlation (24% bracketing), weak negative correlation (58%), and strong negative correlation (90%). A supplemental condition included judges with opposing biases (90% bracketing). We predicted that many participants would equate the accuracy of averaging judgments with the accuracy of the average individual judge. We expected, however, that people would increasingly recognize the effectiveness of averaging as the bracketing rate increased.

Method

Participants. Participants were 145 masters of business administration (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 fifty 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 mean absolute deviation (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 under-estimated 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). Assuming normally distributed errors, averaging outperforms both forecasters in all cases, and by a larger margin at higher bracketing rates. However, 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 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*)

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. 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 misestimate the effect of averaging as the bracketing rate increased across the four conditions ($\gamma = -.31$, $sd = .13$, $p < .01$, 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. Interestingly, Table 1 also reveals that participants who expected averaging to outperform the average judge expected it to outperform both judges in these cases. Few participants gave intermediate estimates. In future research, it would be interesting to test whether this accurate minority is reasoning in a statistically sophisticated way—and therefore sensitive to bracketing information—or is applying a simpler rule that averaging is always beneficial.

In contrast to their estimates for 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 and modal values that participants gave in all conditions. The accuracy of the *alternating* predictions indicates that participants' mistaken estimates for averaging were not the result of the lazy use of "5" as a focal answer, but that they did attempt to calculate the effect of the strategy. (Across conditions, the median participant also estimated that the *confidence* (median = 4.7) and *discussion* (median = 4.92) strategies would outperform averaging.)

By presenting participants with *aggregate* 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 would be applied with concrete instances.

Experiment 2

Experiment 1 showed that, for summary data, most people fail to apply the *abstract* principle that averaging outperforms the average judge. Perhaps people would do better 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.

We expected, however, that the ability to recognize the effect of averaging would depend on an important environmental variable: Are judges observed in isolation or together? Environment is critical because bracketing—which gives averaging its power—is an inherently *dyadic* property. Because people commonly interpret behavior as reflecting a characteristic inherent to an *individual* (Ross & Nisbett, 1991; at least in Western cultures, Morris & Peng, 1992), one would expect observers to attend to individual performance and to translate it to individual ability regardless of environment. But bracketing is nearly impossible to observe 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*. Experiment 2 included two formats, 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.

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. We expected that participants would recognize the benefits of averaging in the simultaneous format, where bracketing is transparent. However, in the sequential format, we expected participants to tend not to appreciate averaging, except in one special case: Opposing biases. Because bias is easily encoded at the individual level ("he tends to underestimate" or "she tends to be an optimist") the bracketing that comes from opposing biases can be reconstructed even when judges are observed in isolation ("an optimist and a pessimist will tend to offset each other's excesses"). Thus, we expected the effect of averaging opposing biases to be recognized regardless of information format.

Method

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

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 ten 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. 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 errors, where Ty had a better MAD than did Chris. Two conditions were used, one in which the Ty was 50% more accurate than Chris (MADs of 31 and 47, respectively) and one in which Ty was 20% more accurate than Chris (MADs of 34 and 42). The MAD manipulation produced no main effects or interactions for the main dependent variable, and will not be considered further. In all cases, averaging produced a MAD lower than either individual's MAD.

Three bracketing rate conditions were created by varying the patterns of error: No Correlation (40% rate), Negative Correlation (80% rate), and Opposing Biases (80% rate). The stimuli for the three bracketing rates are shown in Figure 2 conditions (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).

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*.

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, the error rate was significantly higher in the sequential format than in the simultaneous format condition for the No Correlation condition (.57 vs. .33, n = 87, p < .02, Fisher's Exact test) and for the Negative Correlation condition (.60 vs. .15, n = 87, p < .001, Fisher's Exact test). Also, as expected, the error rate was uniformly low for the Opposing Biases condition, and did not differ by format (sequential .21 vs. simultaneous .19, n = 89, *ns*). As in Experiment 1, participants were less likely to make the error in the Negative Correlation condition than in the No Correlation condition, but only in the simultaneous format condition, where the degree of correlation was transparent (.15 vs. .33, n = 90, p < .03, Fisher's Exact test). Finally, as in Experiment 1, 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 format effect, however, did not hold for opposing biases, which could be encoded at the individual level even in the sequential format. Experiment 2 showed that specific information environments and forms of bracketing helped people recognize the effect of averaging. In the General Discussion, we consider whether these helpful circumstances are common.

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. This is the worst possible performance of 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). Although participants frequently mispredicted that averaging would perform no better than the average judge, certain factors mitigated this mistake: Participants reasoned more accurately when the bracketing rate was high (Experiment 1) and much more accurately when bracketing was made transparent (Experiment 2), either because judges were observed simultaneously or because they had opposing biases.

Unfortunately, there are reasons to believe that most of the facilitating conditions we identified might be uncommon in day-to-day 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 thousands of 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 inter-analyst error correlations or bracketing rates, are typically not provided. Second, even when multiple 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 realized outcome for multiple games. 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—perhaps more so in an individualistic culture—but rarely reveals *interpersonal* patterns of error.

One circumstance in which participants in both experiments recognized the benefits of averaging was when judges exhibited opposing biases, which may be a fairly common pattern in day-to-day life. 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. We should note, however, that the need for averaging is not nearly as essential in this case as in the others: Once the sign and *magnitude* of a given source's bias has been learned, it can be adjusted for without combining across sources (e.g., by subtracting 3 pounds from the bathroom scale reading). We might also note that biases may not always be obvious. For example, 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.

Averaging is a powerful way to reduce error across many settings: Combining the opinions of experts (Clemen, 1989; Hogarth, 1978), integrating the judgments of group members (Einhorn et al., 1977; Gigone & Hastie, 1997), and revising one's own opinion (Soll & Larrick, 2003). 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 simply "locks in" 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 circumstances.

References

Armstrong, J. S. (2001). Combining Forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. New York: Kluwer.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559-609.

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

Bruce, R. S. (1935). Group judgments in the fields of lifted weights and visual discrimination. *Journal of Psychology, 1*, 117-121.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*, 158-172.

Eysenck, H. J. (1939). The validity of judgments as a function of number of judges. *Journal of Experimental Psychology, 25*, 650-654.

Gigone, D. & Hastie, R. (1997). Proper analysis of group judgments. *Psychological Bulletin, 121*, 149-167.

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology, 3*, 398-400.

Gordon, K. (1935). Further observations on group judgments of lifted weights. *Journal of Psychology, 1*, 105-115.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior & Human Decision Processes, 21*, 40-46.

Kelley, T. L. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. *Journal of Educational Psychology, 16*, 300-303.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*, 296-310.

Knight, H. C. (1921). *A comparison of the reliability of group and individual judgments*. MA thesis, Columbia University.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin, 55*, 337-372.

Morris, M. W. & Peng, K. (1992). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality & Social Psychology, 67*, 949-971.

Preston, M. G. (1938). Note on the reliability and validity of the group judgment. *Journal of Experimental Psychology, 22*, 462-71.

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

Soll, J. B. & Larrick, R. P. (2003). *Strategies for Revising Judgment: How, and How Well, Do People Use Others' Opinions?* Manuscript submitted for publication.

Smith, M. (1931). Group judgments in the field of personality traits. *Journal of Experimental Psychology, 14*, 562-565.

Steiner, I. D. (1972). *Group processes and productivity*. New York: Academic Press.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15*, 550-562.

Wallsten, T. S., Budescu, D. V., Erev, I. & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*, 243-268.

Watson, G. B. (1928). Do groups think more efficiently than individuals? *Journal of Abnormal and Social Psychology, 23*, 328-336.

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

Author Note

Richard P. Larrick, Fuqua School of Business, Duke University.

Jack B. Soll, INSEAD.

Correspondence regarding this manuscript may be sent to Rick Larrick at larrick@duke.edu or Jack Soll at jack.soll@insead.edu.

Table 1

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

	Proportion Estimating that Averaging Performs…			
Error Pattern (Bracketing Rate)	No Better than the Average Judge	Better than the Average Judge but no Better than Ms. A	Better than Both Judges	*n*
Positive r (24%)	.74	.00	.26	35
Weak Neg. r (58%)	.55	.05	.40	38
Strong Neg. r (90%)	.47	.00	.53	30
Opposing Bias (90%)	.53	.02	.45	42

Note. Bracketing rate for each condition is given in parentheses.

Figure Captions

Figure 1. Four conditions of error patterns between Ms. A and Ms. B in Experiment 1.

Figure 2. Example of stimuli for the different error conditions in Experiment 2 (Simultaneous Format).

Figure 3. Proportion of participants estimating that averaging would perform no better than the average judge, by error pattern and format.

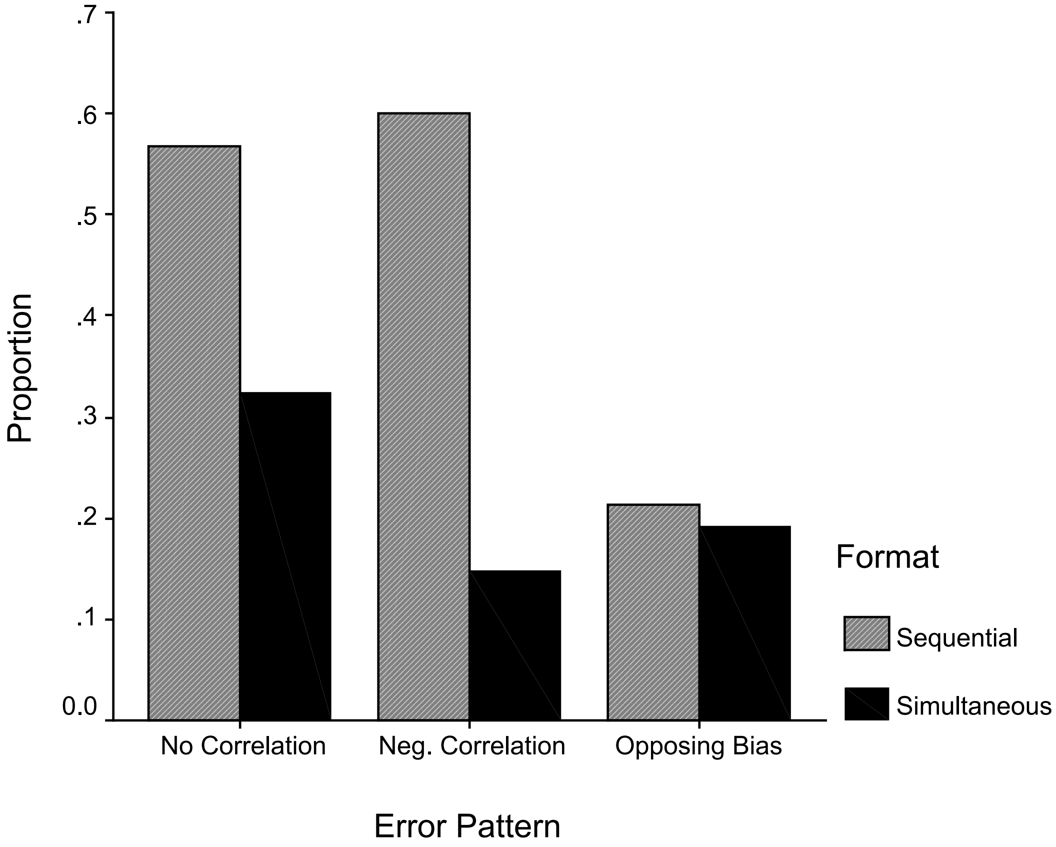
Strong Pos. r		**Ms. B**	
		Over Estimate	Under estimate
Ms. A	Over estimate	18	4
	Under estimate	8	20

Weak Neg. r		**Ms. B**	
		Over Estimate	Under estimate
Ms. A	Over estimate	10	13
	Under estimate	16	11

Strong Neg. r		**Ms. B**	
		Over estimate	Under estimate
Ms. A	Over estimate	3	23
	Under estimate	22	2

Opposing Bias		**Ms. B**	
		Over estimate	Under estimate
Ms. A	Over estimate	2	41
	Under estimate	4	3

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


.7
.6
.5
.4
.3
.2
.1
0.0
Proportion
No Correlation
Neg. Correlation
Opposing Bias
Error Pattern
Format
Sequential
Simultaneous