ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 69, No. 3, March, pp. 237–249, 1997
ARTICLE NO. OB972685

Weighting and Trimming: Heuristics for Aggregating Judgments under Uncertainty

ILAN YANIV

Hebrew University, Jerusalem, Israel

In making major decisions (e.g., about medical treatment, acceptance of manuscripts for publication, or investment), decision makers frequently poll the opinions and subjective estimates of other judges. The aggregation of these opinions is often beset by difficulties. First, decision makers often encounter conflicting subjective estimates. Second, estimates are often expressed with a measure of uncertainty. The decision maker thus needs to reconcile inconsistencies among judgmental estimates and determine their influence on the overall aggregate judgment. In the empirical studies, I examine the idea that weighting and trimming are two important heuristics in the aggregation of opinions under uncertainty. The results from these studies are contrasted with the findings of a normative study using a computer simulation that was designed to assess the objective effects of weighting and trimming operations on the accuracy of estimation. © 1997 Academic Press

It is common practice to poll the opinions and judgments of knowledgeable individuals before making major decisions. Patients often seek several opinions before deciding on radical surgery, faculty deans poll opinions about candidates, journal editors consult referees concerning publication decisions, and business managers seek expert forecasts before embarking on risky projects.

The research presented here concerns two important issues that affect the aggregation process. First, decision makers often encounter conflicting opinions and judgments. Second, judgments are often expressed with a measure of uncertainty. The decision maker thus needs to reconcile inconsistencies among judgmental estimates and determine their influence on the overall aggregate judgment. In the following section, I suggest that weighting and trimming are two important heuristics in the aggregation of opinions under uncertainty. In the studies that follow, I examine the use of these cognitive heuristics in studies in which individual decision makers were required to create aggregate estimates on the basis of different samples of judgments. The results from these studies are contrasted with the findings of a normative study using a computer simulation that was designed to assess the objective effects of weighting and trimming operations on the accuracy of estimation.

This research was supported by grants from the Israel Science Foundation and the Israel Foundations Trustees. I am grateful to Dean Foster for numerous productive discussions on this work and to Josh Klayman and anonymous reviewers for helpful comments on the manuscript.

Address correspondence and reprint requests to Ilan Yaniv, Department of Psychology, Hebrew University, Mt. Scopus, Jerusalem, Israel. Fax: 9722-588-1159. E-mail: msilan@pluto.mscc.huji.ac.il.

AGGREGATION PROCESSES

In the following section, I review the cognitive psychological bases for weighting and trimming processes in judgmental aggregation of estimates. In parallel, I also consider the normative conditions under which one might expect weighting and trimming operations to increase accuracy over simple averaging.

Previous research has considered simple averaging as a psychological model of the aggregation process (Anderson, 1981; Dawes, 1979; Einhorn & Hogarth, 1975; Einhorn, Hogarth, & Klempner, 1977; Hastie, 1986; Sniezek & Henry, 1989). Numerous empirical studies on forecasting and estimation have also suggested simple averaging of individual opinions as a *normative* scheme and a method for improving the accuracy of predictions (Armstrong, 1985; Ashton, 1986; Hill, 1982; Hogarth, 1978; Zajonc, 1962; Zarnowitz, 1984). Due to its prominence in the judgment literature, simple averaging is used as a baseline in evaluating the distinctive effects of weighting and trimming operations.

Cognitive and Normative Bases of Weighting and Trimming

In this study, weighting was evaluated in two contexts. First, I examined the use of weighting as a cognitive heuristic in aggregation. Second, I examined its

0749-5978/97 $25.00
Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.

effect on estimation accuracy. There are several psychological bases for weighting input judgments by the confidence expressed by the judges. Confident judgments are useful for decision makers because they are more informative (i.e., less vague) and hence more conducive to action. Moreover, according to conversational norms and the cooperative principle (Grice, 1975), confidence in judgment is an indicator of the judge's belief in his or her knowledge and hence useful to the decision maker.

From a normative viewpoint, one might ask whether confidence is indeed a valid predictor of accuracy. If confidence is positively correlated with accuracy (Yaniv, Yates, & Smith, 1991), then weighting judgments by confidence would increase the accuracy of the aggregate judgment. Numerous studies indeed have found a small to moderate positive correlation between confidence and accuracy (Armstrong, 1985, p. 138; Braun & Yaniv, 1992; Lichtenstein, Fischhoff, & Phillips, 1982; Wells & Murray, 1984; Winkler, 1971; Yates, 1990). At least one study has found that greater accuracy can be achieved by using confidence as a weighting factor in combining human predictions with statistical, base-rate predictions (Yaniv & Hogarth, 1993).

Another aggregation heuristic considered here is *trimming of outlying judgments*. An input judgment is called "outlying" if it is extreme relative to other opinions in the sample. In some cases we may say that a judgment is outlying if it is not only extreme, but also highly confident. From a cognitive psychological viewpoint, trimming is a simplifying heuristic. First, it reduces the conflict among inputs and thus helps abstract the central tendency of the judgment set. Second, and in a different vein, decision makers can use trimming as a heuristic rule for curbing the influence of individuals who strategically aim to bias the aggregate opinion by announcing extreme judgments with great confidence. Trimming is a two-edged sword, however. In trimming outlying forecasts, decision makers could unknowingly be ignoring their best data—although dissenting estimates differ from the consensus, they are not necessarily wrong. Moreover, a tendency to resolve inconsistencies by trimming outlying opinions, as in discounting evidence that challenges one's prior beliefs, can hamper proper revision of beliefs (Bochner & Insko, 1966; Lord, Ross, & Lepper, 1979).

Statistical analysis delineates some of the conditions under which trimming outliers is a useful strategy in aggregating judgments. DeGroot (1986, pp. 564–569) shows that for samples drawn from a heavy-tailed, symmetric distribution, the trimmed mean has advantages over the sample mean (see also Wilcox, 1992). Because the utility of trimming depends on the parent distribution of the judgments, I will consider next the distributional properties of judgmental errors. The following section also provides the background for our methods.

Properties of Judgmental Interval Estimates

This work is focused on *interval estimates* of uncertain quantities. Intervals are often communicated in the course of real life forecasting and decision making situations. For instance, an expert asked to forecast inflation might produce a finely grained estimates such as "4 to 5%" or a coarser estimation such as "1 to 12%". The width (or graininess) of an interval estimate presumably reflects the individual's assessment of his or her knowledge (Yaniv & Foster, 1995, 1997). Two notable findings on interval estimates are relevant to our work on aggregation of judgments. The first finding arises from numerous studies that have asked subjects to generate interval estimates for uncertain quantities such as general knowledge questions (e.g., "air distance between New York and Chicago"). In such studies typically about 40 to 60% of interval judgments include the true answers (Lichtenstein *et al.*, 1982; Yates, 1990, Chapter 4; Russo & Schoemaker, 1992). This is the case even when subjects are specifically asked to generate interval estimates which include the truth with a probability of 98% (Alpert & Raiffa, 1982). In one previous study that asked subjects to estimate 95% confidence intervals, we observed a 43% hit rate (Yaniv & Foster, 1997). Thus subjective 95% confidence intervals are far too narrow relative to the expected hit rate. Nevertheless, confidence might still be a useful tool for aggregating interval judgments, as shown later.

A second notable finding concerns the extent of disagreement among judges and its consequences for an intersection rule for aggregation. In theory, decision makers can compute the intersection of the set of given interval estimates and then pick a point estimate in that range. In practice, the intersection rule for aggregation is not feasible because the intersection most often does not exist. For instance, I checked the overlap among random pairs of interval judgments drawn from the pool of subjective 95% confidence intervals obtained in Yaniv and Foster (1997). Two intervals overlap if they have at least one point in common. In only 56% of the pairs was there an overlap between the two. In the remaining 44%, the two intervals had not a single point in common. This result is based on 250 samples of pairs of interval judgments drawn with replacement. In another simulation, samples of size $n = 8$ were randomly drawn and the intersection of all eight estimates in each sample was evaluated. Here less than 1% of the

samples of eight intervals had an overlap. In general, as sample size rose from 2 to 8, the chances of an overlap diminished dramatically. The prevalence of no-overlap indicates that the intersection scheme is not a feasible rule for aggregation due to the great disagreement among judges. This underscores the cognitive difficulty that decision makers might encounter in aggregating others' opinions, and hence, the need for aggregation heuristics.

Bases for Weighting Interval Judgments

In the present study, the precision or "graininess" of an interval judgment is viewed as an indication of a judge's faith in his or her knowledge (Yaniv & Foster, 1995, 1997). Graininess thus has a communicative function and decision makers could weight interval judgments by assigning them weights proportional to $1/g$, where g is the precision (interval width) of the judgment.

From a normative point of view, weighting judgments by their inverse width is useful if the width of interval judgments is monotonically related to the magnitude of their errors. Relevant evidence comes from a study by Yaniv and Foster (1997) in which respondents gave subjective 95% confidence intervals for quantities such as "number of countries in the United Nations" or "height of Mount Everest." Only 43% of the confidence intervals contained the correct answers, indicating that subjective 95% confidence intervals are narrower than they should be.

More importantly, however, further analyses revealed that interval width predicts subjects' absolute errors. Absolute error is defined as $|t - m|$, where t is the true answer and m is the midpoint of the subject's interval. In one analysis, for each *individual* a correlation was computed between g (interval width) and $|t - m|$ (absolute error). These individual correlations averaged 0.76. This analysis shows that, for a given individual subject, interval width is a moderate predictor of his or her absolute error. In a second analysis, for each *question* a correlation was computed between g and $|t - m|$. The correlations calculated in this fashion averaged 0.34. Thus, for a given question, subjects who indicate narrower intervals tend to have lower absolute errors. The latter findings imply that weighting judgments by inverse width might improve the accuracy of aggregate judgments.

Bases for Trimming

The results of the 95% confidence-interval study (Yaniv & Foster, 1997) suggest that trimming outlying opinions might be beneficial. Trimming means that people discount input judgments that lie far from the consensus of the sample. As noted, removal of outlier opinions from a sample could be a two-edged sword. Statistically, the usefulness of trimming outliers depends on the parent distribution (DeGroot, 1986, pp. 564–569; Wilcox, 1992). The distribution of judgmental errors for the 95% confidence-interval study (Yaniv & Foster, 1997) is presented in Fig. 1.

The measure of accuracy used is the *normalized error*, defined as $(t - m)/g$, where t is the true answer, m is the midpoint, and g is interval width (Yaniv & Foster, 1997). The normalized error measure implies that, in evaluating the accuracy of an estimate, listeners consider not only the error in the estimate but also the precision claimed by the judge. It captures the intuition that an erroneous judgment stated with great precision (i.e., high confidence) is disliked more than a similar error stated with less precision. For instance, consider the accuracy of two hypothetical judgmental estimates concerning "the number of United Nation members (in 1987)": (A) "130 to 150" and (B) "130 to 132." Although both estimates miss the truth (there were 159 UN members in 1987), they might be evaluated differently. The width of the first estimate is 20, thus its normalized error is less than one unit; in contrast, the second estimate (width = 2) has a normalized error of about 14 units. In terms of normalized error alone, B is less accurate, reflecting the fact that its absolute error is large relative to the level of the claimed precision.

Figure 1 shows a histogram of normalized errors (rounded to the nearest integer) in the 95% confidence-interval study. The bar above zero represents interval judgments that contain the truth (intervals that contain the truth have $-0.5 \leq$ normalized error ≤ 0.5). Although judgments included the correct answer in slightly less than half of the cases, most (86%) of the normalized errors were between -4 and $+4$. The tails represent those estimates that are both far from the

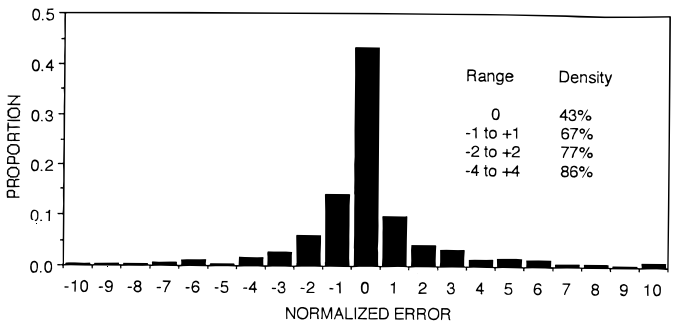

FIG. 1. Distribution of normalized errors from the 95% confidence-interval study (Yaniv & Foster, 1997). Extreme normalized errors (less than -10 or greater than $+10$) occurred in 5.5% of the cases.

correct answer *and* narrow. For illustration, an error-to-precision of 10 would be exhibited by a judge who, in estimating a historical date, stated an interval period of 10 years whose midpoint was off the truth by a hundred years. Only 5.5% of the estimates had absolute normalized errors greater than 10.

The distribution of errors has important implications for the aggregation of opinions. First, it is centered around zero. This means that the central tendency of a random sample is likely to be correctly centered near the true answer. Second, it has relatively thick tails (e.g., by comparison to the normal distribution).[1] Roughly speaking, the distribution is similar in appearance to the Cauchey distribution. This shape of distribution of normalized errors was found in a number of studies using different respondent populations and different general knowledge estimation questions (e.g., Schul & Yaniv, in press; Yaniv & Foster, 1997). According to DeGroot (1986, pp. 564–569), for symmetric distributions with relatively thick tails a *trimmed sample mean* is preferred to the sample mean as an estimator of the central tendency of the distribution. The distribution in Fig. 1 implies that even a small sample of judgmental estimates (e.g., $n = 5$) is likely to include outliers from the tails of this distribution. In sum, these findings suggest that trimming might increase the accuracy of estimation in aggregating judgments.

HEURISTIC AGGREGATION OF OPINIONS

The preceding discussion suggests that weighting and trimming might be two important cognitive heuristics in aggregation of opinions under uncertainty. In four studies I examined how respondents form their best point estimates based on a sample of estimates. In Studies 1 and 2 respondents created a series of aggregate estimates that were each based on a sample of two judgments; the focus of the analysis was on the role of weighting in aggregation. In Studies 3 and 4 respondents were supposed to aggregate on each trial a sample of 5 to 8 judgments. Here the larger samples allowed evaluation of the role of trimming. In each study, the fit of weighting and trimming schemes to subjective aggregate judgments was assessed in comparison to the fit of simple averaging. The latter served as a baseline due to its prominence as a model for the aggregation of forecasts.

In the second part of this article, the results from these four studies were contrasted with the findings of a normative study based on a computer simulation that assessed the actual effects of weighting and trimming operations on accuracy of estimation.

Study 1

The first study assessed the role of weighting in aggregation. Subjects were supposed to form aggregate judgments on the basis of samples of interval judgments. The interval judgments were embedded in a fictional scenario with instructions as follows. “Imagine that you have been chosen to teach for several months in another country. While traveling to your destination you meet groups of students from your home country. You use these encounters to learn about the culture, society, geography, and history of the place. In each of the following cases, we ask you to imagine that you are approaching two individuals for answers to a specific question that you have. Each of the individuals answers your question by providing an estimated range based on his/her memory and best judgment. Your task is to determine what you think the true answer might be based on the two range estimates.” Two sample questions from Study 1 in which subjects were supposed to indicate their best estimates and ranges, are shown below.

In local currency, what is the price of a guided tour through the capital city?

Person A says:	4–100
Person B says:	15–35
Your best estimate	_____
Best range _____	

How many restaurants are in the biggest city?

Person C says:	50–100
Person D says:	110–160
Your best estimate	_____
Best range_____	

The interval estimates presented as the opinions of persons A, B, C, and D were actually sampled from pools of answers to *real* questions collected in an earlier study (Yaniv & Foster, 1997) on judgmental estimation. The two intervals (4–100 and 15–35) used in the present study with the question on “the price of a guided tour” were randomly drawn from the pool of answers given by subjects in an earlier study as estimates of the “number of American symphony orchestras.” Similarly, the ranges given along with the question on “restaurants in the biggest city” were randomly drawn from the pool

[1] A simple informal comparison shows a striking difference between the observed distribution in Fig. 1 and the normal distribution. First note that in the standard normal distribution, the range defined by the z values -0.68 and $+0.68$ contains 50% of the density; similarly, with the observed distribution in Fig. 1, the interquartile range -0.6 to $+0.9$ contains 50% of the distribution. Now in the standard normal distribution the range -2.0 to $+2.0$ contains over 95% of the density whereas in Fig. 1, the range -2 to $+2$ contains only 77% of the observations. Clearly a considerable portion of this distribution lies at the tails (compared with the normal distribution). As a matter of fact, even the range -10 to $+10$ contains only 96% of the density.

of estimates originally made for the "number of stories in Sears tower in Chicago."

The sampling of estimates from pools of real answers is central to the methodology of this research, for reasons that will be explained later. In this study, actual estimates were presented along with fictitious questions to help ensure that respondents focus on the input estimates provided to them and minimize the influence of prior knowledge.

Each subject made aggregate judgment for each of 25 different pairs of intervals. These pairs were randomly sampled from 25 pools of answers to general knowledge questions, as described above. Overall we used 250 samples (pairs) of intervals, divided into ten different questionnaire versions. The subjects were 50 undergraduate students who were randomly assigned one of the 10 versions.

Aggregation Schemes

The analysis evaluates the effect of a weighting operation in subjective aggregation of inputs. Aggregate estimates derived from formal schemes were fitted to the subjective aggregate judgments. The schemes for aggregation were designed to explain how respondents arrived at the aggregate point estimates.[2]

Let $i = 1, \ldots, n$ be the indices of a series of n interval judgments, and let x_i and g_i be the respective midpoint and width of interval i. With a *simple averaging* scheme, equal weights are assigned to all inputs (regardless of interval width). Thus the resulting estimate is the mean of the midpoints of all input intervals,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Simple averaging has been suggested as a model of human judgment. It is also a prominent formal method for aggregating forecasts. I therefore use it as a baseline for comparing other schemes.

[2] The question of how respondents arrive at the aggregate *widths* is not being addressed by the schemes. Note that with respect to *point estimates* the respondent's goal is simply to be as close as possible to the true answer. One can readily evaluate whether respondents engage schemes that increase the accuracy of their point estimates. In contrast, there are different, possibly opposing goals that could (and should) guide respondents in constructing their *best ranges*. Respondents may wish to give intervals that are likely to include the true answers and hence are *fairly wide*. At the same time, respondents may also aim to give intervals that are *not too wide* otherwise they would be uninformative. (This is the accuracy-informativeness tradeoff described in Yaniv & Foster, 1995.) In sum, due to these countervailing objectives the criteria for evaluating the goodness of fit of intervals are more complex and require a separate theoretical treatment.

A second scheme, called *weighting*, implies that respondents assign weights to the input judgments as a function of their width. Specifically, a weighted average of the midpoints of all input judgments x_i is computed with weights inversely related to width,

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i,$$

where

$$w_i = \frac{\frac{1}{g_i}}{\sum_{k=1}^{n} \frac{1}{g_k}};$$

thus,

$$\sum_{i=1}^{n} w_i = 1.$$

Schemes that involve trimming operations are not meaningful with samples of size $n = 2$. Trimming is therefore defined later in Studies 3 and 4, which involved larger samples.

Results

For each pair of input judgments, a weighted average and a simple average were calculated. The resulting statistical aggregate estimates were fitted to the subjective aggregate judgments. The fit of a particular scheme was assessed by computing a normalized error measure $|a - s|/g$, where s is the scheme-derived, aggregate estimate for a given sample, a is the subjective estimate for that sample, and g is the width of the subject's estimated interval. The normalization by g (see also Yaniv & Foster, 1995, 1997) makes it possible to pool the data across questions.

The fits of the weighting and simple averaging schemes to respondents' estimates are shown in Table 1. For each respondent, the mean fit of simple averaging and the mean fit of weighting were obtained. A paired t test indicated that weighting provided a better fit of subjects' estimates than simple averaging, $t(49) = 4.74$, $r < .001$. This effect means that weighting provides a better account of the aggregation process than does simple averaging.

This result is not specific to the error normalization used above. The same conclusion is obtained in using the "absolute percent error" which equals $100 * |a - s|/a$. This alternative method for normalizing the errors has been prominent in the judgment and forecasting

TABLE 1

Fits of Various Schemes to Subjective Aggregate Judgments

Aggregation schemes	Study 1 ($n = 2$)	Study 3 ($n = 8$)	Study 4	
			$n = 5$ with outlier	$n = 5$ no outlier
Trim and weight	—	0.99*	1.11	0.50
		39%**	27%	22%
Trimming (median)	—	1.14	1.13	0.66
		46%	27%	21%
Weighting	0.75	1.12	1.62	0.50
	23%	45%	53%	22%
Simple averaging	2.05	5.18	1.48	1.26
	66%	328%	47%	49%

*This fit measure is the normalized error $|a - s|/g$, where a is the subject's aggregate estimate, s is the statistical aggregate estimate, and g is the width of the subject's interval estimate. Lower numbers indicate a better fit.

**This fit measure is the "absolute percent error" $100 * |a - s|/a$, where a is the subject's aggregate estimate and s is the statistical aggregate estimate. Lower percentages indicate a better fit.

literature (Armstrong, 1985). The mean absolute percent errors (Table 1) were 66 and 23% for the simple averaging and weighting schemes, respectively, $t(49) = 3.98$, $p < .001$.

Study 2

In Study 2 we sought more direct evidence that people assign greater weight to the more precise estimates. We created pairs of questions in which the estimates had the same midpoints, but varied only in their precision. Consider the following matched pair of questions:

Sample 1

What is the number of students that attend the main university in the capital city?

Person A says:	8–24 (in thousands)
Person B says:	11–13 (in thousands)
Your best guess	_____

Sample 2

What is the number of students that attend the main university in the capital city?

Person C says:	15–17 (in thousands)
Person D says:	4–20 (in thousands)
Your best guess	_____

The midpoints of the intervals in samples 1 and 2 are identical (A, C = 16 and B, D = 12). But in sample 1, the lower estimate (midpoint = 12) is associated with greater precision whereas in sample 2 the higher estimate (midpoint = 16) is associated with greater precision. Fifteen pairs of questions (like the one above) were prepared. In each pair, samples 1 and 2 had identical midpoints but differed in precision. Two sets of materials were constructed. Each set included the 15 questions using either sample 1 or sample 2. Respondents ($N = 20$) were randomly assigned one of the two sets.

If respondents place greater weight on the more precise estimate, then their best guesses should be larger in sample 2 than in sample 1 above. This prediction was confirmed. For instance, the mean best guesses for samples 1 and 2 (question above) were 12.7 and 15.5, respectively. In 13 out of 15 pairs of questions, the mean best guesses were biased in the direction predicted by the hypothesis that people weight the estimates by their precision (sign test, $p < .05$). This systematic effect of precision on the subjective aggregate estimates in the direction of the more precise estimate provides direct evidence for weighting.

Study 3

In Study 3, respondents formed aggregate judgments based on samples of eight input judgments, as in the example illustrated below. As in Study 1, various schemes were fit to people's estimates. The larger sample size allowed further examination of the use of weighting in aggregation. More importantly, it enables examination of the procedures for trimming outlier estimates.

In what year was the last earthquake?

Person A:	1600–1900
Person B:	1800–1850
Person C:	1800–1970
Person D:	1870–1890
Person E:	1820–1910
Person F:	1931–1932
Person G:	1700–1800
Person H:	1500–1900
Your best estimate	_____

Best range _____

Forty-two questions were created and then divided into three questionnaires. Each questionnaire included 14 questions with different sets of interval judgments. The sets of input intervals were randomly drawn from the pools of answers of Yaniv and Foster (1997), as in Study 1. Thirty subjects participated in the study; 10 were assigned to each version.

Analysis

As in Study 1, two schemes were fit to subjects' aggregate estimates: simple averaging and weighting. The use of larger samples however readily demonstrates a major weakness of the weighting heuristic that lies in its great sensitivity to opinions. For instance, in the question above, opinion F is extreme relative to the

others. Under weighting, opinion F would be assigned a large weight because it is stated with great precision. Therefore, it would strongly bias the resulting aggregate estimate. People may have procedures for dealing with such outliers. The two following schemes were meant to curb the impact of extreme input judgments.

One important scheme of interest called *trimming*, is based simply on the median. The median operation, by definition, trims all opinions in the sample except for the one (or two) middle opinion(s) (depending on the number of opinions). Another scheme of major interest is *trim and weight*. With this scheme, the extreme opinions are trimmed and the remaining ones are weighted according to their precision. The trim and weight procedure for trimming is described here.

For each interval estimate, an "extremity index" is calculated. Extremity is indicated by the measure $|x_i - \bar{x}_w|/g_i$, where x_i and g_i are the midpoint and width of interval i; $\bar{x}_w$ is the weighted average of all judgments in the sample as defined in Study 1:

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i.$$

An interval judgment i is trimmed if and only if $|x_i - \bar{x}_w|/g_i > c$, where c is some cutoff point. Thus an interval is considered outlying if it is both narrow and far from $\bar{x}_w$. With the trim and weight scheme, the aggregate estimate is the weighted average of the estimates left in the sample *after* trimming. The scheme trim of weight captures in a formal way the simple intuition that opinion F stands out as an outlier because it is both "narrow" and "far from the consensus."

Statistical demonstrations of the use of trimming for samples drawn from heavy-tailed distributions (e.g., DeGroot, 1986, pp. 564–569) recommend trimming of 10% to 20% of the data. Accordingly, I have chosen a cutoff level of $c = 4$. This choice was made heuristically, based on Fig. 1, which shows that 14% of the density lies outside the limits -4 to $+4$. The effects of two alternative cutoff points ($c = 2, 8$) were also tested.

Results

As shown in Table 1, the goodness of fit (based on the normalized error used in Study 1) was significantly better for weighting than for simple averaging, $t(29) = 8.39$, $p < .001$. Trimming (median) also provided a better fit than did simple averaging, $t(29) = 9.49$, $p < .001$. Trimming did not differ from weighting, however. Trim and weight ($c = 4$) provided a significantly better fit than weighting alone, $t(29) = 5.26$, $p < .001$, but did not differ significantly from trimming alone. The mean absolute percent errors (see Table 1) lead to similar conclusions. The corresponding significance tests replicated the results above suggesting that trim and weight was better than either trimming (median), $t(29) = 2.72$, $p < .05$, or weighting, $t(29) = 5.43$, $p < .01$; weighting provided better fit than simple averaging, $t(29) = 13.96$, $p < .001$.

Next, I examined the effects of alternative cutoff levels c for the trim and weight scheme. Note that as c increases the chances that an outlier will be found in a sample decreases. In this study, the percentage of samples that included at least one outlier were 86, 55, and 17% for c values of 2, 4, and 8, respectively. The corresponding mean absolute percent errors for trim and weight were 40, 39, and 43%, respectively. The fit results of trim and weight seem fairly robust to changes of the cutoff level; with $c = 8$ the fit of trim and weight approaches the fit of weighting alone because in fewer samples any estimates are being trimmed.

These results are consistent with the hypothesis that in aggregating judgments individuals weigh inputs according to their precision but trim those that are far from the "consensus." Some converging evidence for a trim and weight heuristic comes from the observation that subjects occasionally crossed out one or two of the input intervals provided to them. The deleted input judgments tended to be either exceptionally wide intervals or narrow, extreme intervals. Weighting and trimming operations indeed assign low (or zero) weights to these types of judgments.

In Studies 1 and 3, the effects of trim and weight and of trimming alone were inferred from statistical fits of aggregate judgments. The samples were drawn from the pools of estimates. There was no attempt to control the frequency of outliers and the dispersion of judgments in the various samples; the frequency and occurrence of outliers were representative of the population from which they were drawn. For instance, 55% of the random samples of size 8 included at least one outlier. (Note that the distribution in Fig. 1 implies that outlying judgments ought to be even more frequent in larger samples.)

The advantage of this sampling approach is that it preserves the ecological validity of the research. Thus subjects were faced with sets of estimates that they might have been offered in realistic situations in response to these types of question. We will see in the second part of the investigation that this sampling approach also facilitates the evaluation of the normative accuracy of various schemes. The disadvantage of the sampling approach, however, is in the lack of control over the exact occurrence of outliers in the samples. In the next study a different approach was taken, whereby the occurrence of outliers in the samples was controlled, so that the effects of trimming could be tested directly.

Study 4

The goal of Study 4 was to examine directly the idea that outlying inputs are given relatively low weights in aggregation. The following two changes were introduced. First, in addition to providing point and range estimates, subjects also ranked the input opinions in the order of importance they had assigned to them in aggregation. Second, the number of outliers in each sample was controlled by design. Two types of samples were used, ones that contained an outlier and ones that did not. This made it possible to compare the ranks assigned to the outlier opinions under each condition. Two sample questions are shown below.

In what year was the last major earthquake?

	Opinions	Ranks
Person A:	1890–1920	A. ____
Person B:	1890–1910	B. ____
Person C:	1931–1933	C. ____
Person D:	1880–1890	D. ____
Person E:	1890–1895	E. ____

Your best estimate _____
Best range _____

Now rank the five opinions from 1 through 5 according to the weight you have given them in your estimation (1 = largest weight, 5 = smallest weight).

In what year was the last major earthquake?

	Opinions	Ranks
Person A:	1890–1920	A. ____
Person B:	1890–1910	B. ____
Person C:	1896–1898	C. ____
Person D:	1880–1890	D. ____
Person E:	1890–1895	E. ____

Your best estimate _____
Best range _____

Now rank the five opinions from 1 through 5 according to the weight you have given them in your estimation (1 = largest weight, 5 = smallest weight).

The questions above are identical with one exception: opinion C is an outlier in the first question but not in the second (although it has the same width in both). In constructing the materials for this study, we specifically selected samples of five judgments that included one outlier. An interval judgment with an "extremity index" (see Study 3) greater than 4 was considered an outlier. For each sample with an outlier, an identical matching sample of judgments was created in which the outlier was replaced with a non-outlier. The serial position of the outlier opinion in the sample varied randomly across questions. Twelve pairs of questions were created. Each pair consisted of two versions of the same questions, with the only difference being whether or not an outlier was included. The two versions of each question were assigned to different booklets. The respondents ($N = 36$) were undergraduate students who received payment for their participation. They each completed one booklet.

As in the previous studies, subjects were asked for their best aggregate estimates and best ranges. Then they were asked to rank the five estimates according to the importance weights (highest to lowest) they had assigned them in aggregation. If two (or more) input opinions were equally important, they could assign them the same rank.

Results

The two primary analyses involved the fits of various schemes and the ranks assigned to the estimates.

Fit. The fits of the various schemes are shown in Table 1. In the *non-outlier condition*, trim and weight was effectively identical to weighting alone (because there were no outliers). Trim and weight was better than simple averaging, $t(35) = 3.23$, $p < .05$, and did not significantly differ from trimming. In the *outlier condition*, trim and weight was better than weighting alone, $t(35) = 2.36$, $p < .05$, but did not significantly differ from trimming alone, $t < 1$. Trimming provided a better fit than simple averaging, $t(35) = 2.44$, $p < .05$; likewise, trim and weight provided a better fit than simple averaging, $t(35) = 2.70$, $p < .05$. The poor fit of weighting clearly demonstrates the inadequacy of weighting alone without provisions for trimming of outliers. Weighting, by definition, placed heavy weights on the narrow outliers—the very estimates that subjects tended to discount. A comparison of the fits between trim and weight and weighting alone across the two conditions (Table 1) suggests that people do trim outliers.

Ranking. The first goal of the analysis was to examine the evidence for trimming by comparing the ranks (1 = high to 5 = low) assigned to outliers and non-outliers. Outlying input judgments were assigned lower weights than the matching non-outliers; the mean ranks were 4.3 vs 2.4, respectively, $t(35) = 10.8$, $p < .001$.

The second goal of the analysis was to examine the relationship between the precision of estimates and their ranking. Generally, precision should correspond to ranking except for cases where a highly precise estimate happens to be an outlier in the sample. We conducted a multiple linear regression analysis where ranking was regressed on two predictors: *precision* and *outlier indicator*. The precision variable was defined as the logarithm of g/g_0, where g is the interval width and

g_0 the median width of the five estimates in the sample. Note that the magnitudes of the estimates varied for the various questions. The scaling by g_0 permitted analyses across all questions. The outlier indicator is a zero/one variable that indicates whether a judgment is a non-outlier (0) or an outlier (1) as they were defined in the design of the study. The regression standardized coefficients for precision and the outlier indicator were 0.33 and 0.53, respectively, with t values of 8.3 and 13.2, $p < .001$. The multiple correlation (R) for this regression model was 0.33 ($R^2 = 0.11$), $F(2, 1437) = 90.6$, $p < .001$.

Overall Study 4 provides three kinds of results that tie together well. First is the comparison of the fit measures across the outlier and no-outlier conditions which suggests that trimming operation is involved in aggregation. Second is comparison of the ranks for the outlier and matched non-outlier estimates. And third are the regression results which suggest that an interval's precision and extremity index are significant predictors of the weight it receives in aggregation. Together these results provide consistent evidence for weighting and trimming operations in aggregation.

CONCLUSIONS

The four studies suggest that both weighting and trimming may be used in a contingent fashion in the aggregation of judgments. With minimal samples of size 2 (Study 1), weighting significantly improved the fit of people's aggregate estimates in comparison with the simple average. With larger samples of size 5 or 8 (Study 3 and 4), weighting and trimming operations consistently improved the fit over weighting alone. Interestingly, trimming (median) alone also provided a better fit relative to weighting. In Study 4, we obtained some direct evidence for weighting and trimming operations.

The conclusions that emerge from the analyses of aggregate estimates naturally lead to the second central issue of this research, namely, the marginal contribution of weighting and trimming procedures to the accuracy of estimation.

NORMATIVE ACCURACY OF AGGREGATION SCHEMES

Suppose a decision maker adheres to particular aggregation heuristic. How accurate would he or she be in the long-run in comparison with someone else who adheres to another heuristic? The goal of the following normative study is to assess the relative accuracy and variability of various schemes. The results of the study provide normative "benchmarks" for the usefulness of strategies that we fitted to people's aggregate estimates. A computer simulation was used to perform the two primary tasks in this study: (i) iterative sampling of opinions from the pools of interval estimates and (ii) calculation of measures of accuracy and variability.

Simulation Study

A computer program was built to simulate the long-run performance of various aggregation operations. The simulation draws random samples from pools of answers to real questions and then produces for each an aggregate estimate of the true answer. Suppose, for instance, that the simulation draws two intervals "4–100" and "15–35" from the pool of estimates made about the "number of American symphony orchestras" (Yaniv & Foster, 1997). The simulation then calculates an aggregate estimate for each scheme. For example, simple averaging of the estimates in the sample above yields 38.5, whereas weighting yields 29.7. The various aggregate estimates are then fitted to the correct answer, which happens to be 31 for this particular question. Because several samples are drawn from each pool, the variability of the resulting aggregate estimates could also be calculated. The process is then repeated for the pool of answers concerning the next question (e.g., "height of Mount Everest"). The fit and variability measures across all pools of answers are then assessed for significance.

Method

This study consisted of 15 large simulation runs. In each run, the sample size was fixed at either $n = 2$, 5, or 8. Also, in each run, eight samples were drawn with replacement from each of the 42 pools of answers (i.e., a total of 336 samples per run). For each sample, statistical estimates were calculated according to each of the aggregation schemes. The mean and variance of these statistical aggregate estimates were calculated across the samples from each pool of answers. As a measure of fit we used, as in Studies 1, 3, and 4, the normalized error $|t - s|/\bar{g}$ where t is the true answer to the question, s is the mean of the statistical aggregate estimates across samples for a given question, and $\bar{g}$ the median interval width in the pool of the interval estimates made about the corresponding question.

Results

The results shown in Table 2 are the average fits obtained from the simulation runs. To assess significance, a binomial test was used. Pairwise comparisons among the various schemes were made using the 42 questions as replicates (recall that each question defined a pool of answers). A particular scheme was deemed better than another scheme if it yielded a better fit on a significant number of pools. With $N = 42$ pools,

TABLE 2

Normative Study: Fits of Various Schemes to the True Answers

Aggregation schemes	Sample size		
	$n = 2$	$n = 5$	$n = 8$
Trim and weight	—	*0.99	0.98
		**43%	39%
Trimming (median)	—	0.97	0.90
		51%	39%
Weighting	1.20	1.17	1.23
	56%	46%	43%
Simple averaging	2.18	2.46	1.83
	151%	223%	140%

*The fit is measured in terms of the normalized error $|t - s|/\bar{g}$, where t is the true answer, s is a statistical aggregate estimate based on a particular scheme, and $\bar{g}$ the median interval width, based on the data from Yaniv & Foster (1997). Lower numbers indicate a better fit.

**This fit measure is the "absolute percent error" 100 * $|t - s|/t$, where t is the true answer and s is the statistical aggregate estimate. Lower percentages indicate a better fit.

the critical value of N^+ is 27, $p < .05$, one-tail, by sign test. Thus, a scheme is better than another if it has a lower fit on (at least) 27 out of 42 pools of intervals.

Accuracy of aggregate estimates. The relevant comparisons among the schemes in Table 2 are within column, separately for each sample size. Consistent with the fit measure shown in the table, weighting was better than simple averaging (for sample sizes of 2, 5 and 8, the N^+ values were 26, 27, and 27, respectively). The trimming operation (median) performed better than simple averaging on larger samples (for samples of sizes 5 and 8, the N^+ values were 28 and 29, respectively). Similarly, trim and weight was better than simple averaging (for samples of sizes 5 and 8, the N^+ values were 36 and 30, respectively). Trim and weight also outperformed weighting alone (for sample sizes of 5 and 8, the N^+ values were 30 and 29, respectively).

These results indicate that weighting and trimming procedures for aggregating samples of judgmental estimates generally improve the accuracy in the estimation of true answers. The usefulness of these operations is contingent on the size of the sample of input judgments. First, with minimal sample size of 2, weighting provides an improvement over simple averaging. With larger samples, either trimming alone or trim and weight improve estimation above and beyond weighting or simple averaging.

Variability of aggregate estimates. Another formal criterion for evaluating aggregation schemes is the variability of the aggregate estimates that are produced by a particular scheme across different samples. Naturally, if two schemes are equally accurate, one might prefer the one that has the lower variability. Variability was measured in terms of the standard deviation of the aggregate estimates across samples from a given pool. The standard deviations were then averaged across all pools of answers, as shown in Table 3. (For purposes of scaling, the standard deviations were normalized by the median interval width in that pool.)

Pairwise comparisons were made among the schemes. As noted earlier, the 42 questions defined 42 pools of answers. By the sign test with $N = 42$, the critical N^+ value is 27 ($p < .05$, one tail) which means that a scheme is significantly better than another if it produces lower standard deviations on 27 or more of the pools (out of 42). Using this test, weighting was better than simple averaging for samples of 2, 5 or 8 estimates, the respective N^+ were 30, 33, and 33. Trim and weight was better than weighting for sample sizes 5 and 8; the respective N^+ values were 27 and 27. Trimming was better than simple averaging for sample sizes 5 and 8, the N^+ values were 33 and 35, respectively. The standard deviations for trim and weight did not differ from those for trimming alone.

The simulation results show that schemes that produce more accurate estimates also tend to have lower variability. In particular, with samples of size 2, weighting dominates simple averaging because it produces higher accuracy and lower variability. With samples of size 5 and 8, trimming (median) as well as trim and weight dominate other schemes, as they have higher accuracy and lower variability.

GENERAL DISCUSSION

Heuristics are ubiquitous in judgment and reasoning processes as they provide approximate, useful solutions to frequently occurring problems. Nevertheless, they lead sometimes to serious, systematic errors through injudicious use (Kahneman, Slovic, & Tversky, 1982). This observation is conducive to two divergent research strategies. Researchers could investigate the "valid scope" of heuristics in attempt to delineate their usefulness across different situations to assess their generality. An alternative research strategy is to explore the

TABLE 3

Normative Study: Standard Deviations of Estimates of Various Statistical Schemes

Aggregation schemes	Sample size		
	$n = 2$	$n = 5$	$n = 8$
Trim and weight	—	0.94	0.72
Trimming (median)	—	1.07	0.66
Weighting	1.86	1.13	0.95
Simple averaging	4.42	3.64	2.07

"limits" of heuristic thinking and to delineate the areas where it fails. With the latter approach, special questions or tasks are constructed so as to reveal situations where the heuristics yield erroneous or illogical answers. These two approaches might be seen as complementary.

The work presented in this article involved the former research approach. Accordingly the use of aggregation heuristics was considered across a variety of cases and samples. I reported two sets of results: the first was based on empirical studies of decision makers' aggregation heuristics (Studies 1–4); the second came from a computer simulation of aggregation. The general conclusion that emerges from these two sets of results is that people seem to use aggregation schemes that are warranted by the simulation results.

The rationale for using a computer simulation of aggregation was that the usefulness of any aggregation scheme ultimately depends on the properties of the judgments that are being aggregated (e.g., the variability of opinions, bias, and frequency of extreme opinions). The present aggregation simulation sampled from pools of judgments that were obtained from respondents who participated in an earlier study (Yaniv & Foster, 1997). In that respect,the samples of estimates were ecologically representative of the estimates that might be obtained in daily life when seeking answers to questions. The results of the aggregation simulation show that weighting and trimming operations improve the accuracy of aggregate estimates over simple averaging. Whereas with small samples of opinions (e.g., 2 opinions) weighting improves accuracy, with larger samples (e.g., 5 to 16), trimming alone as well as the scheme called "trim and weight" outperform simple averaging. It appears that the usage of weighting and trimming operations contingent on sample size dominates simple averaging.

The results of the computer simulation presented here establish therefore norms that could be used as a benchmark for assessing human behavior. In Studies 1–4, decision makers formed a global opinion based on the opinions of several individuals. Conceivably there is an infinite number of potential schemes that decision makers could use for aggregation. It is practically impossible to examine (or even specify) any great number of them. This work has, therefore, been limited to examination of a number of basic aggregation operations that are likely to be the "building blocks" of more complex schemes. The fit results for judges' aggregate estimates suggest that weighting and trimming operations are likely to be components of the subjective aggregation scheme. For instance, based on the fit results one can readily reject the hypothesis that people average all opinions without regard to their extremity in the sample. The relative fit results for various schemes (Studies 1, 3, and 4) are consistent with the interpretation that decision makers (a) generally weight opinions according to their precision in samples of 2 or more opinions and (b) trim extreme opinions in samples of 5 or 8. The results should be viewed as suggestions for the operations that are involved in aggregating judgments.

Additional evidence supporting this conclusion comes from Study 2 which provides direct evidence that in weighting two opinions, decision makers bias their aggregate estimates in the direction of the more precise estimate. Also, Study 4 reveals that the ranking of the input opinions in the sample ($N = 5$) is directly related to their precision, with the exception of the extreme opinions (which are assigned low ranks even when stated with precision). The pattern of ranking revealed in this study is consistent with the trim and weight scheme which implies that an opinion is assigned a low (or zero) weight either when it is stated coarsely or when it is both precise and extreme. The results suggest that cognitive heuristics involving weighting and trimming operations are likely to play a role in people's aggregation of opinions.

It should be noted in passing that the present schemes can be elaborated on in various ways. For instance, the weighting factor was proportional to the inverse of width ($1/g$). Alternatively, weighting could be proportional to $1/\log g$ or $1/g^k$ ($k > 0$). Several weighting systems have been explored in the course of data analysis and were not reported because they yielded indistinguishable results from those obtained in weighting by $1/g$. Similarly changing the cutoff level for the trim and weight scheme (from 2 to 8) had no qualitative effect on our conclusions. I suggest that we are witnessing here another case where human judgment is highly robust to variations in modeling, as has been shown numerous times in the literature on linear models of human judgment (Dawes, 1979; Einhorn & Hogarth, 1975). The conclusion is that it might be sufficient to consider the simpler instances of each family of descriptive models.

Normative Implications

A notable finding from the simulation results is that a few judgments are sufficient to obtain most of what there is to be gained from aggregation. For instance, the accuracy of aggregate estimates increases only to a moderate degree as sample size varies from 2 to 8 opinions. This conclusion holds for weighting, trimming and for trim and weight. Several investigators have already noted that the majority of the gain from averaging large numbers of opinions can be obtained by aggregating as few as two to five opinions (Ashton & Ashton,

1985; Libby & Blashfield, 1978; Hogarth, 1978). At first glance, it seems puzzling that additional data (i.e., opinions) do not boost accuracy to a greater extent. Moreover the relative success of the trimming (median) procedure—which by definition trims all observations except for the middle opinion(s)—suggests that the magnitudes of extreme opinions are dispensable. The following discussion provides an intuitive explanation of why that might be the case.

Note that in forecasting, in general, the goal of taking a sample of opinions is to estimate some external, logically independent quantity (e.g., tomorrow's temperature). Hence, the sample mean does not necessarily approach the true parameter in a lawful manner as dictated by the law of large numbers. In particular, consider the population of opinions about a given question and call the "mean judgment in the population" m and the "true answer" t. The law of large numbers implies that the mean of a random sample of opinions tends to m as sample size increases. If $m = t$, then accuracy should improve as sample size increases. However, if the mean opinion is biased relative to the truth ($m \neq t$) then aggregation can boost accuracy only up to the difference between t and m. Suppose the underlying bias $|m - t|$ is small relative to the standard deviation of the opinions about their mean m—in this case the gain from aggregation is expected to be substantial because averaging will converge on a value close to the truth; in contrast, when the bias is large relative to the standard deviation, the expected gain in accuracy from aggregation is small (Einhorn *et al.*, 1977).

The relative magnitudes of bias and standard deviation might vary across different questions within a study. Whereas for some questions the gain is substantial for others it might be minute. This perhaps explains the common findings that aggregation of opinions tends to improve accuracy but that the gain accrued diminishes rapidly as sample size increases. In general, the bases for weighting and trimming are contingent on the properties of the parent distribution of the judgmental errors. For instance, whereas trimming greatly contributes to accuracy when the distribution of errors is heavy-tailed and symmetric, its effect on accuracy could be lower with skewed or asymmetric distribution and even disappear in the case of a symmetric thin-tailed distribution of errors. It is conceivable also that the advantage of trimming relative to simple averaging would be somewhat lower when the estimates are made on a bounded scale (e.g., a Likert 7-point rating scale or a percentage scale).

In sum, the usefulness of aggregation of judgments is an empirical possibility that depends on properties of the estimates, thereby highlighting the need for computer simulations as a tool for normative study of aggregation schemes.

Cognitive and Social Issues

Whereas the normative justification for trimming depends on the distribution of judgments about the true answer, the cognitive reasons for trimming are more diverse. One could attribute to the decision makers in our study a rational understanding of the distributional properties of the judgmental errors such as the thickness of its tails and prevalence of outliers. An alternative possibility is that decision makers are not cognizant of the properties of the inputs. Instead they simply engage a basic cognitive process that trims data as a part of a generalized strategy of resolving inconsistencies and conflicts among input opinions by removing the dissonant data, whether justified or not. Whereas removing inconsistent data is an anathema for researchers, it appears to be a rational heuristic in intuitive aggregation of judgments given the distribution of errors.

The only basis for weighting opinions in this work was the precision of the input estimates (construed as a proxy for confidence, Yaniv & Foster, 1995, 1997). Because precision is to some degree correlated with error, the weighting scheme improved accuracy over simple averaging. It might be noted that realistic environments are sometimes richer and provide various kinds of information about the judges. Thus decision makers might have available additional cues on which to base their weighting of the judgments, such as the judges' expertise and the judges' reputation based on their prior performance. The possibility of weighting each opinion by multiple factors presents an interesting theoretical problem; it also suggests a complicated cognitive process for the decision makers who construct the weights (e.g., see Klayman, 1988, on the learning of cues in a probabilistic environment).

CONCLUSION

The aggregation of judgments under uncertainty is a complex cognitive task yet a practical problem that occurs in many decision making situations. Facing complex tasks, individuals often rely on heuristics that provide approximate solutions. The results of the studies here suggest that people rely on weighting and trimming heuristics. The computer-simulated competition suggests that these heuristics are indeed justifiable on the grounds that they increase the accuracy of estimation.

REFERENCES

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Ashton, R. H. (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes*, **38,** 405–414.

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

Armstrong, J. S. (1985). *Long-range forecasting: From crystal ball to computer* (2nd ed.). New York: Wiley.

Braun, P. A., & Yaniv, I. (1992). A case study of expert judgment: Economists' probabilities versus base-rate model forecasts. *Journal of Behavioral Decision Making*, **5,** 217–231.

Bochner, S., & Insko, C. A. (1966). Communicator discrepancy, source credibility, and opinion change. *Journal of Personality and Social Psychology*, **4,** 614–621.

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

DeGroot, M. H. (1986). *Probability and statistics*. (2nd ed.) Reading, MA: Addison-Wesley.

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

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Decision research* (pp. 129–157). Greenwich, CT: JAI Press.

Hill, G. W. (1982). Group versus individual performance: Are $N + 1$ heads better than one? *Psychological Bulletin*, **91,** 517–539.

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

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14,** 317–330.

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

Lichtenstein, S., Fischhoff, B., & Phillips, P. (1982). Calibration of probabilities: The state of the art of 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). New York: Cambridge University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, **37,** 2098–2109.

Russo, J. E., & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, **33,** 7–17.

Schul, Y., & Yaniv, I. (in press). Inferring accuracy for judges and items: Choice of unit of analysis reverses the conclusions. *Journal of Behavioral Decision Making.*

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

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives*. New York: Cambridge University Press.

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science*, **1,** 101–105.

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

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

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

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

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, **110,** 611–617.

Yates, J. F. (1990). *Judgment and decision making* (pp. 75–111). Englewood Cliffs, NJ: Prentice Hall.

Zajonc, R. B. (1962). A note on group judgments and group size. *Human Relations*, **15,** 177–180.

Zarnowitz, V. (1984). The accuracy of individual and group forecasts from business and outlook surveys. *Journal of Forecasting*, **3,** 11–26.

Received: August 9, 1996