ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 43, 1–28 (1989)

Accuracy and Confidence in Group Judgment

JANET A. SNIEZEK AND REBECCA A. HENRY

University of Illinois at Urbana–Champaign

A weighting scheme model is used to describe and evaluate the process by which groups combine individual judgments and their associated confidence levels into a single group judgment with some level of confidence. In general, the group judgment process is not well described as an averaging process. Group judgments were, with few exceptions, significantly more accurate than mean or median individual judgments, leading to a 23.7% reduction in standardized bias. Further, 30% of the group judgments were *more* accurate than the group's *most* accurate individual judgment. Two factors related to increased accuracy through grouping were (a) high disagreement, i.e., large variance, in initial judgments, and (b) group judgments *outside* the range of initial individual judgments. In addition to being more accurate, groups were generally more confident than individuals. Ratings of confidence and accuracy, but not 99% confidence interval size, were correlated with actual judgment accuracy. Limitations of the weighting scheme approach are discussed, with suggestions for further research on group judgments under uncertainty.

Accurate human judgment of unknown quantities has great utility in many organizational contexts, but can be difficult to achieve. A common practice is to assign multiple persons to the judgment task. But while the use of multiple judges may increase confidence in judgments, it will not necessarily increase accuracy. The general purpose of this paper is to describe and evaluate group judgments and their associated confidence assessments.

There are many ways in which a single final judgment can be obtained from a set of judges, but here we restrict attention to the most common method, the group meeting. Using the properties defined by Steiner (1972), the task in such meetings can be characterized as (a) *unitary* in that division of labor according to subtasks is not feasible, (b) *discretionary* because the task does not constrain the group to a particular set of procedures for combining individual contributions, and (c) *optimizing* because accurate judgment is the desired output.

We thank James H. Davis, Fritz Drasgow, Reid Hastie, Robin M. Hogarth, Patrick Laughlin, Joe McGrath, J. Edward Russo, Paul Shoemaker, Ira Solomon, and an anonymous reviewer for useful conversations and comments on an early version of this paper. Address all correspondence, including reprint requests, to Professor Janet A. Sniezek, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820.

0749-5978/88 $3.00

But group judgment tasks have an additional unique characteristic. By definition, judgments of unknown quantities will always have some degree of uncertainty associated with them. Thus, unlike with some other unitary, optimizing, and discretionary tasks (e.g., "eureka problems"), all group members in a judgment task can be expected to have at least some *uncertainty* about their answers.[1] The kinds of tasks that are frequently assigned to groups are those in which it is difficult to know when a high quality outcome has been reached (Hart, 1985). Judgment *accuracy,* defined as the proximity of a judgment to the actual criterion value, cannot be determined at the time of the judgment. Therefore the quality of a judgment must itself be estimated—a judgment of a judgment called a *confidence assessment.*

CONFIDENCE IN JUDGMENT

It is the level of uncertainty (or conversely confidence) in a judgment that typically determines if and how it is used in organizational decision making. In practice, *some* indication of the uncertainty surrounding a judgment is necessary in order to use the judgment at all. Thus it can be just as important to determine *how* and *how well* individuals and groups assess their uncertainty as it is to determine *how* and *how well* individuals and groups form their judgments. A sample of confidence assessments can be described in terms of the level of confidence, and evaluated on two dimensions, calibration and validity. In this paper, the samples of individual and group confidence assessments are compared to each other in terms of level of confidence, calibration, and validity.

Calibration

Calibration reflects the appropriateness of the expressed level of confidence, i.e., the match between obtained accuracy and expressed confidence level over a large set of responses. There is much evidence that individuals are not well calibrated in the assessment of their responses in nontrivial tasks, but are prone to overconfidence (cf. Einhorn & Hogarth, 1977; Lichtenstein, Fischhoff, & Phillips, 1982). That is, one judges one's responses to be more accurate than they really are. For example, in a calibrated sample of 90% confidence intervals around judgments (see Seaver, von Winterfeldt, & Edwards, 1978), 90% of the actual criterion values should be contained between the upper and lower limits of the intervals. Instead, fewer than 90% of persons' "90% confidence intervals" typically contain the criterion value.

[1] If one can be perfectly certain about the accuracy of one's response, then the task should be regarded as a knowledge task, not a judgment task. Similarly, if uncertainty about the actual value of the criterion is *not at all* reduced following the production of a response, then that response is defined as a *guess,* not a judgment.

Just as with individual confidence, it is useful to know the appropriateness of group confidence assessments. Individual and group calibration in confidence assessments are compared in this study. Increased confidence (Castore & Murnighan, 1978) or commitment (Maier, 1967) in group output can be desirable objectives of group meetings, but overconfidence can also be an undesirable symptom of groupthink (Janis, 1972). Tversky and Kahneman (1974) attribute the overconfidence observed with individual confidence intervals to a process of anchoring on the judgment and not adjusting sufficiently in setting the limits of the confidence interval. However, a group will have k judgments to serve as k anchors and k confidence intervals that will expand the range of values deemed possible by one or more group members. For this reason it seems unlikely that groups will be subject to the same degree of overconfidence bias due to anchoring with insufficient adjustment. Thus it is predicted that groups will be less overconfident than individuals.

Validity

If one can determine the relative quality of one's multiple judgments, then there will be a positive relationship between judgment accuracy and confidence across multiple items. Such a relationship would be evidence of the *intraindividual validity* of confidence assessments. In a group setting with multiple judges and a single item, the analogous relationship between judgment accuracy and confidence demonstrates *interindividual validity*. There is clearly an advantage to having some indication of the *relative* quality of a set of judgments, even if confidence assessments are poorly calibrated. In practice it may even be more important to know *which* are the better of a set of judgments than to know the accuracy of the set as a whole. We assume that judgment accuracy and confidence *should* in a normative sense, be correlated dimensions of a response.[2] But, in practice, there will be various valid as well as invalid cues to judgment accuracy that could be used to form confidence assessments. The important empirical question that we seek to answer is whether the confidence assessments do demonstrate interindividual or intergroup validity.

WEIGHTING SCHEMES IN GROUP JUDGMENT

In this paper, judges' individual and group judgments and their confidence assessments are compared. More specifically, group judgments and confidence assessments are described and evaluated in terms of relevant *weighting scheme* models. A simplified representation of the group judgment process is given in Fig. 1. The group judgment process is described

[2] Thus judgment and confidence are distinguished from other uses of two response dimensions, e.g., *preference* and *confidence* as defined by Stasser and Davis (1981), that are presumed to have no necessary dependence.

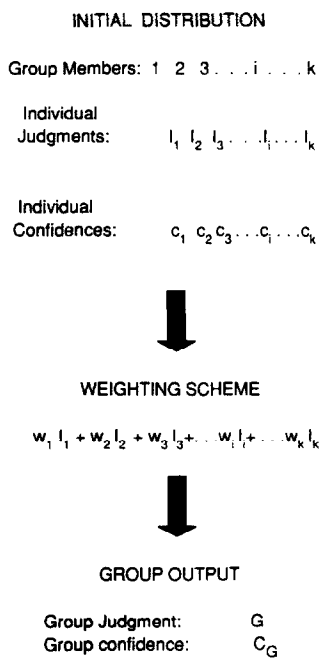

FIG. 1. A representation of the group judgment process.

by a weighting scheme for combining k individual judgments, with k associated confidence levels, into consensus group output. In an attempt to describe *how* groups form judgments, the relative fits of four weighting scheme models to the group judgment process are compared. To determine *how well* groups form judgments and confidence assessments, the quality of group judgment and confidence assessments are compared to that of individuals, and to judgment and confidence assessments from applications of relevant weighting schemes.

The weighting scheme approach permits us to describe the relation between initial state and group performance. The group judgment process is modeled as a weighting scheme for combining k individual judgments into a single group judgment. Let I_i be the i^{th} individual's judgment of the value of a criterion variable Y. Given k individual judgments, $I_1, I_2, \ldots, I_k$, the group judgment G is given by $G = \Sigma\, w_i I_i$ for $i = 1$ to k, where $w_1, w_2, \ldots, w_k$ are the weights assigned to the k individual judgments, and $\Sigma\, w_i = 1$.

The observed accuracy of group judgments relative to individual judgments will depend on several factors, including the independence of errors in individual judgments, number of judges (Hogarth, 1978), and amount of *bias* in individual judgments (Einhorn, Hogarth, & Klempner,

1977). Bias in the individual judgments is given by $B = \mu_I - Y$ where μ_I is the mean individual judgment and Y is the actual criterion. As will be explained in the forthcoming pages, the amount of bias in a set of individual judgments is of particular interest because it determines whether the group should use an averaging or a nonaveraging process to obtain an accurate group judgment. Once the k individual judgments have been formed, group judgment accuracy will depend on the manner in which the k weights are assigned. To evaluate the quality of actual group judgment, we compare observed performance with the performance of a set of three weighting schemes: *equal, median,* and *best* models. These models differ with respect to the rule used to form a group judgment from k individual judgments.

Equal Weighting

The *Equal* weighting scheme produces a group judgment by weighting each individual judgment equally: $G_M = \Sigma\, w_i I_i$ where $w_i = {}^1/k$ for all i. Let μ_I be the population mean of individual judgments and Y be the actual criterion value. If individual judgments are unbiased and errors $e_i = I_i - Y$ are independent, the expected error with the mean model, $E(G_M - Y)$ is zero. Under these circumstances, the Equal judgment will generally be more accurate than most individual judgments. Thus, little improvement in accuracy can be gained by using group meetings if individual judgments are unbiased. Statistical averaging of individual judgments with the equal model will always be faster than equal weighting in a meeting. And, statistical averaging will prevent groups from using some basis for variable weighting, e.g., status, talkativeness, or experience, that may be actually invalid. While it is often true that the Equal weighting scheme will be more accurate than behavioral aggregation with unequal weights, it must be kept in mind that the improvement in accuracy will be lessened to the extent that individual judgments are biased (Einhorn *et al.*, 1977), as many estimates in nontrivial tasks are likely to be.

Because averaging of individual judgments will enhance accuracy only if the individual judgments are unbiased, nonaveraging processes are needed in many, if not most, group judgment situations. The real need for a process beyond simple averaging exists in situations where there are reasons to suspect nonzero bias, i.e., $E(G_M - Y) \neq 0$, and reasons to want to avoid bias. These are the situations in which human judgment and group meetings are commonly used. Unlike with low bias, equal weighting is not more accurate than variable weighting if bias exists. If group judgment is characterized by an Equal weighting scheme regardless of bias, then group judgments will be no more accurate than statistical averaging.

Although the process of combining individual responses in group deci-

sion making has been assumed to be an averaging process (Davis, 1973; Graesser, 1982), the nature of the process that groups use to form a judgment is not known. Yet, as can be seen from the foregoing discussion, group judgment accuracy is dependent on the *match* between group judgment process and bias. Thus, it is of great interest to learn whether groups use averaging or nonaveraging processes under conditions of bias in the group members' individual judgments.

One potential influence on the manner in which the group forms a consensus judgment is the initial distribution of individual judgments. Distributions of individual judgments will vary with respect to many features, but the effects of two features, bias and disagreement, on the process used by the group to form a group judgment are of particular interest. The mean of the distribution of individual judgments determines the amount of bias in the k persons' initial judgments, and can be interpreted as an index of group members' ability level. Since any weighting of k accurate judgments will produce a more accurate group judgment than any weighting of k inaccurate judgments, mean individual judgment bias is expected to be inversely related to group judgment accuracy. Rohrbaugh (1979) found accuracy of initial judgment policies to be a significant predictor of group policy quality.

Disagreement, given by the variance of the k individual judgments, is expected to influence the weighting scheme used to form the group judgment. The context in which the group operates will have at least as much information available as an individual, and perhaps much more (Maier, 1967). Uniqueness of information, true and false, manifests itself in the variance of the individual judgments for a particular item. If most people share the same information (or biases), their judgments will be similar and the set of judgments will have a small variance. In this case of low disagreement, grouping is unlikely to change the similar judgments; social information provides consensual validation for the individual judgments. Low disagreement should provide each group member with external confirmation of the quality of his or her judgment, thereby reducing the apparent need for further discussion and evaluation of the relative accuracy of each others' judgments. As a consequence, an Equal weighting scheme will be more likely to be used if disagreement is low, regardless of bias. Therefore, it is predicted that, with low disagreement, judgments produced from group interaction will be no more accurate than the mean individual judgment. With averaging of the individual judgments, groups will yield more accurate judgments only to the extent that random error is reduced. Bias will be relatively unaffected.

If instead, individuals have unique and conflicting information about the criterion variable, there is likely to be "cognitive conflict" (Brehmer, 1976) in the form of discrepant judgments. Cognitive conflict is distin-

guished from motivational conflict in the sense that all parties have the same goal, accurate judgment in this case, but different information (or biases). The adequacy of group vs individual judgment will depend on how a single group judgment is formed from the conflicting individual judgments.

In contrast to low disagreement, high disagreement can be expected to lead to nonaveraging processes. A high variance in individual judgments has been found to be related to the magnitude of the difference between average individual and group judgments (Libby, Trotman, & Zimmer, 1987). Though nonaveraging processes could potentially lead to either greater or lessor accuracy than an Equal weighting scheme, the relevant literature suggests that high disagreement will result in nonaveraging processes that increase group judgment accuracy compared to the Equal weighting scheme. It has been suggested that the effectiveness of groups relative to individuals will be increased by heterogeneity of group members (Steiner, 1972), a wide range of opinion (Lorge, Fox, Davitz, & Brenner, 1958), or low agreement in the form of low interjudge correlations (Hogarth, 1978). Rohrbaugh (1979) reports similarity of initial individual judgment policies to be inversely related to final group judgment quality.

Variable Weighting

The alternative to equal weighting is to vary the influence of the k group members' individual judgments on the consensus group judgment. In the general case, $G = \Sigma\, w_i I_i$, with ${\sigma_w}^2 > 0$, and $\Sigma\, w_i = 1$. An extreme case of the variable weighting scheme, the *one-for-all* strategy will be considered first. With the all-for-one strategy, the group selects one member's individual judgment for the group judgment. That is, $w_i = 1$ for individual i and $w_j = 0$ for all $i \neq j$. But how does the group choose this member? For now, two selection procedures are considered: the selection of the individual with the *median* judgment in the distribution or with the actual *best,* i.e., most accurate, judgment. Later we shall look at selection rules based on individual confidence.

The median rule. The median rule selects the median individual judgment for the group judgment.

$$G_C = I_{(k+1)/2} \qquad \text{when } k \text{ is odd,}$$

where $I_1, I_2, \ldots, I_k$ are the order statistics. This model will yield results approximate to those of the Equal model as long as distributions of individual judgments are symmetric about the mean. The symmetry assumption will be likely to hold if $B = 0$. For $B > 0$, however, distributions of individual judgments may be markedly skewed, with large differences

between the median and mean judgments. The median rule reflects a different type of averaging process than the Equal weighting scheme. In essence, the median rule selects the judgment of the "compromise candidate," and completely ignores all extremists. Which of the two averaging strategies actually leads to more accurate judgment will depend on the form of the distribution of I_i. Because of asymmetric individual judgment distributions, both the mean and median provide useful baselines for evaluating group judgment accuracy in this study.

The best rule. Another one-for-all weighting strategy of interest in this study is that of the *best* rule. The best model identifies, once the actual criterion value is known, the maximum performance that could have been obtained from the given set of individual judgments (Einhorn *et al.*, 1977). The best individual judgment, i.e., the one with the shortest distance to the actual criterion value, becomes the group judgment: $G_B = w_i I_i + w_j I_j$ with $w_i = 1$ for min. $|I_i - Y|$ and $w_j = 0$ for all $i \neq j$.

If group judgments are consistently equal to the most accurate of their individual judgments, then we can conclude that groups can evaluate the relative accuracy of their k judgments and identify their best members. Though group performance at the level of the best member has been reported by Einhorn *et al.* (1977) and Uecker (1982), actual group judgment is usually inferior to that of the best member (Hill, 1982; Hastie, 1986; Lorge *et al.*, 1958).

The best rule sets an important standard for evaluating group performance, but it does have limitations. Groups are said to exhibit "process loss" (Steiner, 1972) to the extent that their actual group output is inferior to potential group output, i.e., that of their best member. But in the case of tasks involving judgment under uncertainty, the "best" judgment capitalizes in part on chance. It is also possible to define "process gain" as the extent to which group output is superior to that of the best member. With unbiased judgments, it is possible for an Equal weighting scheme to occasionally produce judgments that are better than the best individual judgment.[3] *Thus, for chance reasons, the Equal weighting scheme applied to samples from a population of unbiased judgments will always demonstrate both process loss and process gain.* But such evidence of process gain due to chance is much *less* likely to occur with biased judgments. When all individual judgments are biased, groups can outperform their best member *only* by adopting a nonweighting strategy and producing a judgment that is *outside* the range of the individual judgments. Einhorn *et al.* (1977) reported that 15% of their groups of size three were more accurate than their best members in a population estimation task with

[3] This will also be true of any variable weight model that assigns weights to two or more group members. Similarly, any all-for-one model will occasionally produce judgments that are as good as the best individual judgment.

standardized bias of .42. Uecker (1982) found the majority of groups to perform at the level of their best member in a sample size estimation task. Unfortunately, it is not known whether the groups in either of these studies achieved their increased accuracy with a group judgment inside or outside the range of the initial individual judgments. One limitation of the weighting scheme approach to modeling the group judgment process is that it does not take into account the possibility that the group judgment may fall outside the range of the initial individual judgments.

The variable weighting scheme. One final weighting scheme of interest is the general case of variable weighting: $G_v = \Sigma\ w_i I_i$, with $\sigma_w^2 > 0$, and $\Sigma\ w_i = 1$. If group judgment is characterized by a *variable* weighting scheme, the accuracy of the group judgments will depend on the rule for assigning weights to individual judgments. As long as the weights are inversely proportional to error, G_v will be superior to G_M. The problem facing the group is to determine the relative accuracy of the group members' judgments so that valid weights can be assigned to them.

In practice, it will be quite difficult for groups to identify and use valid cues to accuracy. Various appealing variables, e.g., status, experience, and talkativeness, may in fact have zero or even negative validity. Equal weighting models have been shown to outperform differential weighting models under many circumstances, partly because there is no danger that weights can be reversed (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). Group judgments will be more accurate than equal weighting if weights are proportionate to accuracy. But how can accuracy be evaluated a priori? As Einhorn *et al.* wrote, ". . . the question remains as to how well groups can identify and weight their better members *before* the true value becomes known" (1977, p. 159). One possibility is to look at group members' assessments of the accuracy of their own judgments, i.e., their confidence assessments.

There are two meaningful questions regarding the role of confidence assessments in the group judgment process. The first concerns whether confidence assessments *should* be used to determine weights for combining individual judgments. This question is answered by the interindividual validity of the confidence assessments. If, in a set of k judgments and confidence assessments, the individual confidence assessments do have *inter*individual validity, then they would provide a valid cue to the *relative* accuracy of the judgments that could be used as a basis for determining weights. With valid confidence assessments, it might even be feasible to replace behavioral aggregation with statistical aggregation. The second question is whether confidence assessments *are* actually used, regardless of their interindividual validity, in the weighting of group members' judgments. This question is answered by the fit of a particular variable weighting model, Confidence, with weights proportionate to confidence to ac-

tual group judgments: $G_c = (\Sigma\ c_i I_i)/\Sigma\ c_i$ where c_i is the reciprocal of the width of the confidence interval.

Effective communication of true individual uncertainty during group meetings is in general quite difficult. Individual differences in *self*-confidence or self-esteem may be great enough to dominate true beliefs about the quality of one's judgment itself. Explicit expressions of individual uncertainty in judgment in the form of confidence intervals should have at least two advantages over ordinary means. First, uncertainty is expressed in terms of the criterion variable, not in terms of oneself. Discrepancies between true and manifest beliefs should be less likely to result from personality variables in statements about the criterion variable than with statements about one's own confidence. Second, confidence intervals allow for unlimited discrimination among levels of uncertainty. In contrast, ordinary verbal and nonverbal means of communicating uncertainty in meetings is extremely imprecise and ambiguous. Thus it is of interest to determine whether explicit statements about the degree of uncertainty associated with one's individual judgments are or can be used as a basis for determining weights that improve group judgment accuracy compared to equal weighting. The search for unequal weighting aggregation rules to improve judgment accuracy has a long history, but improvement over equal weighting is typically slight at best (Ferrell, 1985). In this paper, the feasibility of using confidence intervals in aggregation rules is explored by first investigating their interindividual validity.

To achieve the objective of evaluating group judgment and confidence as outlined above, it is desirable to obtain judgments under conditions of varying bias. Because previous work in the area of risk judgment has detected various biases in individual judgment, this area appeared to be a suitable one for empirical investigation of group judgment and confidence. The observation that little is known about social factors in risk judgment and the belief that risk judgment is itself an important area of study provided additional reasons to use a risk judgment task for the study of group judgment accuracy and confidence. Finally, pilot testing of the task of estimating the annual frequency of deaths from various causes revealed the task to be one in which participants vigorously discussed a wide variety of opinions and information. Relevant studies of individual risk judgment are summarized below.

Risk Judgment

To determine the adequacy of individual judgments of risk, Lichtenstein, Slovic, Fischhoff, Layman, & Combs (1978) obtained estimates of the frequencies of 41 causes of death. Although the judged frequencies correlated highly with actual frequencies, bias in judgments was evident. Rare causes of death were overestimated while common causes of death

were underestimated. But the magnitude and direction of judgment error in individual risk assessment can depend on cognitive factors such as the manner in which judgments are elicited (Fischhoff & MacGregor, 1983) or the availability of anchors (Lichtenstein *et al.*, 1978).

Risk judgment can also be viewed as a social process (Douglas and Wildavsky, 1982). Although many choices under risk are made by individuals (e.g., the choice of wearing a seat belt, smoking cigarettes, or using oral contraceptives), the individual's risk judgments are typically formed in a social context. In addition to objective data (such as statistics regarding side effects from oral contraceptives on package inserts) and opinion of experts (one's physician, for example), the individual person is surrounded by the subjective judgments of others (including friends and family). As noted by Festinger (1954), when objective reality is not available to anchor judgments and beliefs, persons rely on information provided by others. Deutsch & Gerard (1955) referred to the tendency to accept information from another as evidence about reality as "informational social influence." Yet studies of the adequacy of risk judgments have not studied social influence. A better understanding of risk judgment can result from obtaining risk judgments in a context of social interaction.

Group judgments of risk are also of interest because many decisions under risk are made collectively, by advisory committees, grass roots organizations, board of directors, etc. Because multiple experts involved in the decision-making process can have great disagreement about risks (Hammond, 1984), the group judgment process is all the more important to study.

In summary, individuals have appropriate risk judgments in that they can order causes of death according to their true frequency. Magnitude judgments, however, are biased. It is difficult to interpret observed overestimation and underestimation. Such biases could reflect inaccurate risk judgment or effects associated with the manner in which judgments are elicited. The present study uses a within-subjects design, with response format held constant, to obtain judgments from individuals and groups. It will be possible to compare the extent and pattern of bias in the individual vs group samples of judgments.

SUMMARY OF RESEARCH QUESTIONS

This study attempts to describe and evaluate group judgments and their associated confidence assessments. First, the accuracy of group judgments of risk is compared to that of individuals. Then, four weighting scheme models—Equal, Median, Best, and Confidence Weighted—are fit to the actual group judgments to determine their adequacy in describing the group judgment process. It is predicted that with low disagreement groups will use an averaging process; and as a result, be no more accurate

than the Equal model. In contrast, high disagreement is hypothesized to lead to nonaveraging processes that are expected to improve group judgment accuracy relative to the Equal model. The Equal, Median, and Best weighting schemes are also used as baseline models to evaluate observed group judgment performance.

Similarly, the amount of observed group confidence is compared to that of individuals and those obtained with three confidence weighting models: Mean, Median, and Most. The quality of both individual and group confidence assessments is examined in terms of calibration and validity. Individuals are expected to exhibit overconfidence because of the process of anchoring with insufficient adjustment. But because groups will have available multiple anchors and an increased range of values with multiple confidence intervals, they are expected to have lower levels of confidence, i.e., wider confidence intervals. Finally, we examine the appropriateness of participants' postgroup individual ratings of the quality of their group judgments relative to other groups and on an absolute basis.

METHOD

Participants

Students enrolled in an M.B.A. program (n = 54) were recruited for participation through handouts announcing the study. The announcement indicated that half of the participants, selected by performance quality, would earn money in amounts between $5 and $50. Participants were randomly assigned to groups of three. To prevent pretesting conversations or research into public health statistics, the content topic of the task was not disclosed before the study, and all testing was completed in two sessions on the same day. The most accurate group received $50 for each member, the next two groups $25 for each member, the next two groups $10 for each member, and the next four groups received $5 for each member. The most accurate three participants in the remaining nine groups were given $25, $10, and $5; all other participants were not paid. All participants followed all parts of the instructions satisfactorily so no data were omitted from analyses.

Materials

Participants were given two forms, one for individual answers and the other for group answers. The general instructions indicated that the money awards would be dependent on following instructions in the entire task to prevent unequal incentives for individual vs group performance. Each listed the 15 causes of death in a column. To the left of each cause was a blank for the rank of the frequency, to the right was a blank for the frequency of deaths for a U.S. population of 230,000,000. To the right of

this column were two blanks, "lowest" and "highest" for the limits of the 99% confidence intervals around each frequency estimate. Instructions, written on a separate page stapled on top of each form, requested participants to rank order and estimate the annual number of deaths in the U.S., assuming a population of 230,000,000, for each of the 15 causes. For the confidence interval response, instructions were "In the blanks to the right of your estimate, write the lowest and the highest the frequency could actually be. You should be 99% confident that the true number of deaths falls between these two amounts."

A final post-task form was used to obtain ratings, on a seven point numerical scale, of the group as a whole (on cooperation, accuracy of group answers, disagreement about answers, confidence in group answers, and conflict). Finally, participants were asked to use their own confidence in their group's estimates to predict which category their group would be in: top 2%, next 10%, next 12%, next 26%, or bottom 50%. These categories approximated the actual percentages of participants receiving the various levels of awards, from $50 to $0.

Procedure

Data were collected in two separate halves of a large auditorium. One area, with individual seats, was designated the individual work area. The group areas in the other region had large tables with three chairs. After participants arrived at the research location, they were randomly assigned sets of forms. On the cover of each set was an assigned seat number for individual work and an assigned area for group work. To discourage communication among participants, they were widely spaced over each area; a maximum of 25% of the seats were in use at any time. Members of the same group were assigned individual seats as far apart as possible. Three experimenters monitoring the data collection observed no instance of communication between individuals or groups. Participants appeared to be highly involved in the task.

After completing the task individually, participants went, with their completed answer form, to their assigned group areas. When all three members were present, they were instructed to work as a group with the goal of reaching a consensus on the best possible answers.

When the group portion of the task was finished, participants returned to their individual seats to complete the post-task form. All participants finished the three parts, individual task, group task, and post-task rating, in 1.5 hr.

RESULTS

Group vs Individual Judgment

Accuracy of judgments. Differences in the appropriateness of risk judg-

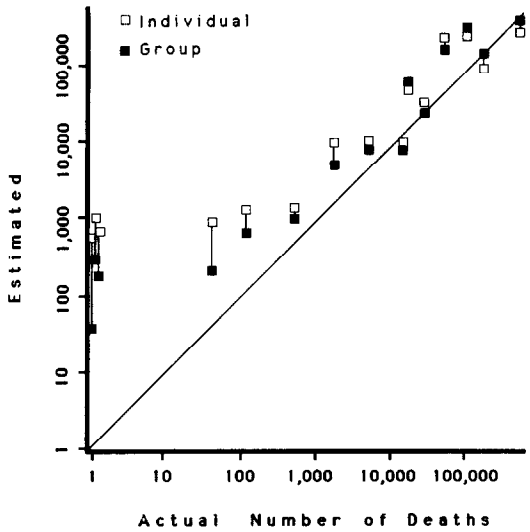

FIG. 2. Geometric means of individual and group judgments over the 15 causes of death.

ment between individuals and groups can be seen by comparing average individual and average group estimates to the actual values (i.e., those values given by public health statistics). Average individual and group judgments are given in Fig. 2.[4] They are expressed in terms of the geometric mean because the distributions were skewed. Accurate estimation of frequency of death would result in data points falling on the diagonal line. The geometric mean of the 18 group judgments is more accurate, i.e., closer to the true value, than the geometric mean of the 54 individual judgments for 12 of the 15 causes of death—all except lung cancer, homocide, and accidental falls. A Wilcoxon rank sum test on total Absolute Percent Error (APE) over the 15 causes of death showed groups to be significantly more accurate than individuals ($W = 2.84$, $p < .003$).

Standardized bias was computed for both the individual and group

[4] Data from Lichtenstein *et al.* (1978) show systematic overestimation of low frequency causes of death by individuals. The more common causes of death are not underestimated as much in the present study; all estimates are larger. As noted by Fischhoff & MacGregor (1982) and Lichtenstein *et al.* (1978), the magnitude of estimates is susceptible to response format and anchoring effects. Unlike in the Lichtenstein *et al.* study, no standard stimulus was used in the present study. And, rather than giving a rate per 100,000 response, participants in the present study gave the total frequency in a population of 230,000,000. This figure or even the estimate for the first cause, lung cancer, could have served as an anchor. Or, perhaps increased attention to the higher frequency causes of death in recent years improved the accuracy of available information or increased apparent size. It will always be difficult to interpret the absolute magnitude of risk judgments; a better approach is to determine what factors lead to relatively more accurate judgment.

judgments for each cause by dividing the bias by the standard deviation of judgments for that cause. Individual standardized bias ranged from −.40 to +.45. Standardized bias was reduced with grouping for each of the 15 causes, with a mean reduction of 23.7%.

Bias, disagreement, and group judgment. The overall effectiveness of grouping for each group was measured by the number of causes for which group judgment was more accurate than the mean individual judgment. The 18 variances of the initial judgments were rank ordered for each cause, and then, to obtain an overall measure of relative disagreement in the groups, the 15 ranks were summed for each group. A Spearman correlation found group effectiveness to be significantly related to disagreement ($r(14) = .52$, $p < .05$). A similar procedure was used to obtain rankings of the 18 groups with respect to overall bias. Group effectiveness was not significantly related to bias ($r(14) = .29$, $p > .10$).

Accuracy of rankings. Judgments of relative risk were evaluated by comparing observed rankings and true rankings of the 15 causes according to frequency. The mean absolute difference between true ranks and individual ranks was 1.68, and between true ranks and group ranks was 1.13. The group ranks were superior to the individual ranks ($t = (70)$, $p < .05$).

Weighting Schemes

Four weighting scheme models were fit to the data to determine their relative adequacy in describing the actual group judgment process: Equal, Median, Best, and Confidence. The number of times that each of the four models provided the best fit for each of the 270 group judgments was: Equal, 27.58; Confidence, 36.58; Median, 64.92; and Best, 140.92. These frequencies differed significantly from the 67.5 frequency expected for each model by chance ($\chi^2 = 117.73$, $df = 3$, $p < .0001$). Table 1 gives values of the root mean square (RMS) of the error between the actual group judgment and group judgment predicted by each of the four models.

APE scores were computed for actual group judgments and those were obtained with each of the equal, median, and best models. Table 2 lists the number of the 18 groups that outperformed each of these models. To get overall measures of judgment quality, mean APE (MAPE) scores over the 15 causes were computed for each group. In terms of MAPE scores, 17 of the 18 actual group judgments were superior to the mean judgments, 16 were superior to the median judgments, and 9 were superior to the best judgment.

The total 270 group judgments (15 causes by 18 groups) were classified with respect to (a) their relation to the range of their respective individual distributions (inside vs outside the range) and (b) their accuracy relative to their best individual judgment (better than best vs poorer than best).

TABLE 1
COMPARISON OF WEIGHTING SCHEME MODELS TO THE GROUP JUDGMENT (RMSE)

	Weighting scheme model			
Cause of death	Equal	Median	Best	Confidence
Heart disease	3,792,942	3,678,565	2,653,769	4,038,387
Stroke	2,396,308	2,369,918	2,653,769	2,378,559
Lung cancer	3,474,697	2,651,692	2,685,483	4,435,016
Motor accident	3,615,214	2,177,675	2,146,331	4,540,488
Breast cancer	1,247,735	1,254,214	1,185,123	1,223,191
Homicide	1,804,002	1,779,656	1,766,073	1,927,425
Accidental falls	429,503	305,717	294,687	522,787
Drowning	223,116	234,878	234,385	253,415
Firearm accident	698,393	707,778	705,917	680,806
Appendicitis	412,434	412,518	412,255	418,543
Syphillis	109,683	128,760	129,224	90,499
Bites and stings	137,923	116,450	116,449	180,110
Whooping cough	78,623	67,896	70,484	85,750
Botulism	442,889	469,529	471,170	378,543
Small Pox	57,458	57,583	58,765	58,124

TABLE 2
NUMBER OF GROUPS OUTPERFORMING MODELS

	Baseline model		
Cause of death	Equal	Median	Best
Heart disease	14	13	7
Stroke	14	11	4
Lung cancer	14	11	6
Motor accident	16	13	3
Breast cancer	13	13	4
Homocide	17	13	8
Accidental falls	15	14	6
Drowning	15	12	5
Firearm accident	13	12	5
Appendicitis	12	11	4
Syphillis	14	13	5
Bites and stings	14	12	7
Whooping cough	17	16	7
Botulism	17	15	4
Smallpox	17	12	5
Mean	14.8	12.7	5.3

Note. Total number of groups = 18.

TABLE 3
CLASSIFICATION OF GROUP JUDGMENTS WITH RESPECT TO RANGE AND BEST

Relation to range	Relation to best individual judgment		
	Better	Poorer	Total
Inside	42	151	193
Outside	39	38	77
Total	81	189	270

Table 3 shows the number and the percentage of group judgments in each of the four combinations. The number of outside group judgments correlated significantly with actual frequency of deaths, Spearman $r(15) = -.83$, $p < .001$.

Group vs Individual Confidence in Judgments

Amount of confidence. Overall, groups set smaller confidence intervals around their group judgments than did individuals. Median confidence limits for groups and individuals over the 15 causes of death are shown in Fig. 3. Because the group and individual confidence intervals have different locations on the log frequency scale, visual comparison of their relative sizes is difficult. Thus it must be noted that, for all causes of death

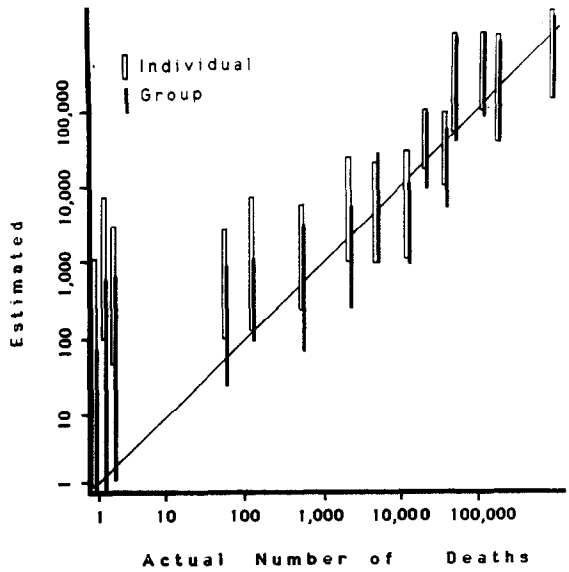

FIG. 3. Median limits of individual and group confidence intervals over the 15 causes of death.

except drowning and lung cancer, the difference between the median upper limit and the median lower limit is smaller for the sample of groups than for the sample of individuals.

To take into account the fact that confidence interval widths depend on the magnitude of the judgment, confidence interval widths were divided by the associated judgments. The resulting measure of confidence is referred to as Proportionate Confidence (PC). PC values for each person and for each group were summed across the 15 causes. A Wilcoxon rank sum text on overall confidence did not show the group median PC to be significantly different from the median PC for their individual members ($W = -.62, p > .05$).

Table 4 shows the number of groups with higher confidence (i.e., smaller confidence intervals) than their average individual confidence, median individual confidence, and greatest individual confidence. Overall confidence for group judgments was higher than the average confidence in 15 groups, than the median confidence in 12 groups, and than the greatest confidence in 4 groups. Two of the groups had overall confidence levels less than the least confident of their members.

Calibration of confidence assessments. Though the confidence intervals set by groups were narrower than those set by individuals, they more often contained the true frequency. Groups were therefore better cali-

TABLE 4
NUMBER OF GROUPS MORE CONFIDENT THAN MODELS

	Baseline model		
Cause of death	Mean	Median	Most
Heart disease	12	11	2
Stroke	14	11	4
Lung cancer	13	11	5
Motor accident	14	10	5
Breast cancer	14	13	6
Homocide	14	13	3
Accident falls	12	11	3
Drowning	11	10	4
Firearm accident	14	12	5
Appendicitis	13	12	5
Syphillis	12	7	1
Bites and stings	12	6	3
Whooping cough	12	10	3
Botulism	15	9	3
Smallpox	17	13	10
Mean	13.3	10.6	4.2

Note. Total number of groups = 18.

brated than individuals for every cause of death except appendicitis. The percentages of individual and group confidence intervals containing the true value for each cause of death are shown in Table 5. While half of the group confidence intervals contained the true value (Mean = 7.5), only one-third of the individual confidence intervals did (Mean = 5.1; $t(70) = 2.27$, $p > .02$).

Relationship of Accuracy to Confidence

Pearson correlations between PC and APE were computed for each cause of death for both the individual and group samples. These correlations were transformed to Fisher z_r scores. The resulting mean z_r values were not significant for either individuals ($r(53) = -.05$, $p > .05$) or groups ($r(17) = +.04$, $p > .05$). However, the sums of group members' ratings of confidence and accuracy were significantly related ($r(18) = .82$, $p < .0001$).

Post-Task Ratings and Performance

Ratings of various features of the group process and group output from the three group members were summed to obtain one rating score for each group. These scores were correlated with actual overall accuracy (APE). Significant correlations were found with rated accuracy ($r(18) = -.60$, $p < .004$), disagreement ($r(18) = -.50$, $p < .02$), and confidence ($r(18) = -.74$, $p < .001$), indicating that groups that actually had higher accuracy

TABLE 5
PERCENTAGES OF GROUP AND INDIVIDUAL 99% CONFIDENCE INTERVALS INCLUDING ACTUAL VALUES

Cause of death	Individuals ($n = 54$)	Groups ($n = 18$)
Heart disease	30%	39%
Stroke	30	44
Lung cancer	15	56
Motor accident	35	56
Breast cancer	33	39
Homocide	30	61
Accidental falls	7	33
Drowning	37	56
Firearm accident	41	56
Appendicitis	41	39
Syphillis	39	50
Bites and stings	31	61
Whooping cough	35	39
Botulism	22	44
Smallpox	37	78

rated their groups to have higher levels of accuracy, confidence, and disagreement.

In their ratings of their group accuracy compared to that of other groups, all but one person estimated that their group would be in the top 50% of groups in terms of overall performance.

DISCUSSION

Group vs Individual Judgment

The first set of findings regarding the relative accuracy of individual and group risk judgments shows the superiority of group judgments. Compared to the sample of individual judgments, the sample of group judgments is more accurately ranked and closer to the actual criterion values. Thus a somewhat different picture of risk judgment emerges with the use of group judgments than with individual judgments. The overestimation and underestimation biases observed with individual judgments are reduced considerably through group interaction. Thus the apparent adequacy of people's risk judgments depends not only on how people are asked (Fischhoff & MacGregor, 1982), but also in which social context they are asked. Because so much of the actual decision making regarding risks is made by groups or individuals in social contexts, future risk judgment research should not be limited to the study of individuals.

The observed 23.7% reduction in standardized bias is double the one-eighth of a standard deviation unit reported by Hastie (1986) to be the customary amount of improvement in judgment accuracy with group interaction. Our analyses reveal two factors related to increased accuracy through grouping: high disagreement and out of range judgments.

Disagreement. As noted by Maier (1967), disagreement in groups can be either an asset or a liability. In this study, groups with relatively larger variances in their distribution of individual judgments showed more improvement over average individual accuracy. This agrees with the findings of Rohrbaugh (1979) and Libby *et al.* (1987). As predicted, disagreement is related to the use of nonaveraging processes to form the group judgment. Without disagreement, it is difficult for group members to distinguish among their judgments to find a basis for variable weighting. The data also show a significant association between absolute levels of group judgment accuracy (APE) and rated disagreement. The more disagreements that group members reported, the more accurate were their group judgments. But in contrast to Rohrbaugh's study, bias was not a significant factor in grouping effectiveness. The negative (but nonsignificant) correlation between bias and grouping effectiveness indicates that initially more accurate groups improved less, perhaps because of a ceiling effect. In summary, disagreement was more important in effecting process gain than was the competence of the individual group members.

Weighting Schemes

The increase in accuracy from grouping is clearly not attributable to mere reduction in random error through averaging. The fact that most group judgments show less bias than the mean of their individual judgments demonstrates that the group process cannot be described by an Equal weighting scheme. Indeed we find that the Equal weighting model provides the poorest fit to the actual group judgments. Similarly, the group process cannot be described by the simple rule of selecting the median value from the distribution of individual judgments. The majority of actual group judgments were superior to median judgments. Therefore there is little evidence of averaging. Averaging, however, is predicted by Steiner (1972) when all members are unsure and have no intense personal commitment to specific alternatives. In the present risk estimation task all participants must necessarily have at least some uncertainty about their judgments. Perhaps a better formulation of Steiner's hypothesis is that groups will average if all members are *equally* unsure and have *equal* levels of commitment to specific alternatives. Yet it is also reasonable to expect averaging if all members are highly certain to achieve, for example, equality of influence or consensus through compromise. Such equality or compromise was definitely not the rule in this study.

This lack of averaging could be considered surprising if it is noted that the formation of a group judgment from multiple individual judgments is conceptually similar to the integration of information from multiple sources by individuals. In this literature, averaging is commonly reported (Shanteau & Nagy, 1982), perhaps because the task is unfamiliar or complex (Hammond, 1980). Yet the data in the present study show the group judgment process to be a nonaveraging strategy.

The relatively good fit of the Best model and the superiority of actual group judgments over the mean and median judgments indicate that the groups formed judgments *as if* unequal partially valid weights were assigned to initial individual judgments. The poor fit of the Confidence model in this study allows us to conclude that groups did not use confidence interval sizes to determine weights. More importantly, the lack of a systematic relationship between accuracy and confidence interval size shows that confidence as it was operationalized in this study is not a valid cue to judgment accuracy and therefore should not be used as a basis for variable weighting.

Out-of-Range Judgments

But it is clear that the group interaction cannot be completely described by means of weighting schemes. The fact that nearly 15% of all group judgments were better than the best judgment and were obtained by going

outside the range of individual judgments demonstrates a large and important increase in accuracy through grouping that exceeds previously reported gains (see Hastie, 1986), and deserves further study. Altogether, 30% of the group judgments were better than best, double the number reported by Einhorn *et al.*, 1977.

At this point, we can identify several plausible reasons for the group accuracy effect that may be used to design future research. The explanations considered do not address the reduction of random error. They refer to increases in accuracy due to either (a) out-of-range judgments, i.e., group judgments more extreme than the most extreme individual judgment, or (b) variable weighting scheme judgments, i.e., group judgments more extreme than only the average individual judgment. Because past studies have not reported out-of-range group judgments, several of the explanations focus on the unique features of the task in the present study: the use of confidence intervals accompanying judgments and multiple items along a single dimension.

First, we consider the "range expansion hypothesis." The use of confidence intervals forces judges to evaluate the quality of their judgments and consider a *range* of alternative values, not just a single value. Thereby, their uncertainty about their judgments is more salient to them and it becomes easier for judges to consider and accept judgments which differ from their initial estimates. Given k confidence intervals and k judgments, the entire range of values available for consideration in group discussion, between the outer limits (of the minimum lower limit to the maximum upper limit) of the k confidence intervals, is broader than it is with only k judgments. Judgments that are outside the range of the k individual judgments can still be within the range of the k confidence intervals. Thus we see how the use of confidence intervals can result in out-of-range group judgments. But to explain the high accuracy of the group judgments, we must assume that groups can identify the relative accuracy of values within the outer limits. If so, then to improve group judgment accuracy, one need only expand the range of values that the group can consider. And, since judges are typically overconfident in that their confidence intervals are too narrow, perhaps group judgment accuracy could be increased even further by enlarging confidence intervals. This might be accomplished by having judges supply reasons that the actual criterion values could be below their lower limit or above their upper limit (see Koriat, Lichtenstein, & Fischhoff, 1980), or simply having judges multiply their confidence interval size by some constant greater than one prior to group discussion.

A second way of looking at the effect is to say that individuals are simply more regressive in their judgments than groups are, i.e., individuals give estimates that are relatively closer to some estimate of the grand average value across all the causes. It is also true that individuals show

greater uncertainty than groups do. It is, of course, quite reasonable to be more regressive if more uncertain. This raises the possibility that groups are not really more able to make accurate judgments than individuals—they only happen to do so as a consequence of being more certain. That is, group interaction reduces uncertainty and therefore regressiveness is reduced. This "uncertainty–regressiveness" hypothesis actually leads to recommendation contrary to those given for the "range expansion" hypothesis: if judges can be made to become *more* confident, they will make more extreme judgments and therefore become more accurate. The social interaction that occurs in the group is one way to increase confidence.

Another possible explanation for judgment shifts in grouping involves the use of a wide range of actual risks in the present task evokes judgments of relative as well as absolute magnitude, and that relative judgments, e.g., a low risk, increase in weight when multiple persons agree to them. Thus the absolute judgment becomes modified in the direction of the relative judgment: the low estimates become lower and the high estimates become higher. Such an "adjustment by relative judgment" effect could be expected to be greater in the present study because judges were required to rank explicitly the causes in terms of their relative frequencies. By using a ranking strategy first, the task becomes one of using item ranks to set group judgments. This is quite different from using the individual judgments for a single item to set a group judgment. Group judgments that are set with respect to a ranking of the cause could go out of the range of individual judgments. Like the uncertainty–regressiveness explanation, this explanation implies that increased accuracy is merely a by-product of other grouping effects. If so, then we can expect to observe increased accuracy with groups only if initial individual bias is such that low frequency items are overestimated while high frequency items are overestimated. Indeed we do observe that out-of-range group judgments are in the direction of lowered estimates for low valued items and higher estimates for high frequency items. Because all low valued items were initially overestimated, the changes result in improved accuracy. In contrast, individual judgments of the high frequency items were not uniformly biased, and out-of-range judgments were often in the wrong direction. But when reduction in standardized bias is considered overall, not for just out-of-range judgments, there is no difference in the amount of improvement between high and low frequency items.

Explanations of polarization shifts from individual to group responses, e.g., informational influence (see Kaplan & Miller, 1983), may also be applied, but it must be kept in mind that the present task differs from typical polarization tasks in that judgment accuracy is the main goal, and judgment accuracy alone was rewarded. Finally, the magnitude of the present finding of increased accuracy through grouping may be attributable simply to increased effort due to the large incentives for accuracy.

Participants in previous studies showing no or small accuracy increases through grouping (cf. Hastie, 1986; Lorge *et al.*, 1958) may have made greater use of satisficing, e.g., averaging, strategies.

Of course, these explanations are not mutually exclusive. Further research is necessary to identify the task features, confidence intervals, multiple items, explicit rankings, large incentives, etc., that lead to high group judgment accuracy, and why. One possibility is that effect is not unique to groups. Such knowledge could potentially be applied to obtain the same high level of accuracy (at lower cost) from individuals. It is also important to determine whether group judgments are generally more accurate or just more extreme than individual judgments. If the latter, then group judgments will not necessarily be more accurate with other patterns of bias in individual judgments, e.g., systematic overestimation of unknown quantities.

One potentially fruitful method of learning about group changes in judgment, whether to increased accuracy, extremity, or confidence, is to observe and describe the social interaction that occurs among group members. The above explanations can be used to develop specific hypotheses about the type and amount of information used in group discussion. For example, the range expansion hypothesis predicts that the group will discuss a wider range of values than that containing the individual judgments.

An apparent limitation of the weighting scheme model is that it cannot describe the group process that results in out-of-range judgments. This problem may be overcome by conceptualizing the group judgment as a two-stage process (Sniezek & Henry, in press). First, information and judgments are exchanged, with ongoing revision of individual judgments. Then, there is the process of weighting of postinteraction revised judgments. While revision is largely an individual process, weighting occurs at the group level. In this manner it might be possible to describe the formation of a group judgment as the averaging of k *revised* individual judgments rather than the unequal weighting of k initial individual judgments. This type of model could then account for both out-of-range and in range judgments.

Confidence in Judgments

As individuals, participants displayed the typical levels of overconfidence with their too-narrow subjective confidence intervals. This particular form of overconfidence has been attributed to a process of anchoring on the judgment and insufficient adjustment from the anchor (Tversky & Kahneman, 1974). Contrary to expectation, groups did not set wider confidence intervals; they were in general even more narrow than those set by individuals. But in groups that have available multiple anchors from the distribution of k individual judgments and confidence intervals, over-

confidence cannot be explained as a bias resulting from anchoring and insufficient adjustment. Both the group and individual confidence intervals displayed extreme overconfidence.

Given that grouping increased judgment accuracy in this study, it is not inappropriate for grouping to also have the effect of increasing confidence. Not only were groups generally more accurate and confident than individuals, they were also better calibrated. But this result is somewhat misleading because better calibration was achieved because of better accuracy, not lower confidence. To see this suppose that we remove the group advantage of better accuracy by adding an amount δ to equate the average individual judgment with the average group judgment. Then the individuals, with their wider confidence intervals, will be better calibrated than the groups. The wider a confidence interval, the more likely it is to contain the true value. In general one can improve calibration, i.e., reduce overconfidence, by simply lowering confidence. These groups, however, achieved better calibration despite narrower confidence intervals. The groups knew more about risk, but they did not necessarily know more about how much they knew. Of course, with only the within subjects design, other explanations, such as an increase in confidence due only to amount of experience in the task (Paese, 1988; Sniezek & Reeves, 1986), cannot be ruled out.

The strongest interpretation of the lack of association between confidence interval size and judgment accuracy is that people have no awareness of the quality of their judgments. To the extent that confidence determines the amount of participation in group interaction (Hastie, 1977) or influence (Johnson & Torcivia, 1967), this raises serious questions about the use of multiple judgments by decision makers when external criteria are not available. It also could partly explain the undemonstrated validity of the Delphi procedure and its popularity (see Sackman, 1974). An implicit assumption in the Delphi procedure is that judges can properly evaluate the accuracy of their own judgments relative to the average group judgment so that they can determine how much to yield to the average group judgment in revising their own judgments (Sniezek, in press). The idea that "those who know should know that they know, and those who don't should know that they don't" is an intuitively appealing idea, but not necessarily true. Winkler (1971) did not find that weighting group members' estimates by their ratings of confidence improved judgment accuracy compared to equal weighting, implying that self-assessments of judgment quality were invalid.

But it is not necessarily the case that confidence assessments have no interjudge validity. In the post-task rating data, significant relationships were found between overall group judgment APE and both rated accuracy and rated confidence. Because these ratings were made by individuals about group judgments, this is a very specific demonstration of the inter-

judge validity of confidence assessments. The lack of agreement between this finding and the nonsignificant relationship between confidence intervals and judgment accuracy for both individuals and groups may be due to the manner in which confidence is assessed—through ratings or confidence intervals. With confidence intervals, there may be large differences between individuals and groups of individuals in the manner in which confidence in judgment becomes translated into limits of the interval. Though interval sizes have been found to be related to confidence ratings (Larson & Reenan, 1979), people may still have more problems expressing their confidence with them than with ratings. If true (as opposed to reported) levels of confidence in judgment are positively related to judgment accuracy, advances in the measurement of the construct "confidence in judgment" would have practical implications for the improvement of judgment through the use of multiple judges. Part of the solution lies in explicit penalty functions for confidence responses that deviate from true judgments (Winkler, 1967). Although the reward system in the present study and the instructions were intended to encourage honest responding, participants might have been more concerned with quality judgments than with quality confidence intervals. Yet, for the reasons noted above, evidence of a correlation between accuracy and confidence intervals across multiple persons may be difficult to obtain.

The final method of obtaining confidence assessments involved participants' individual judgments of the quality of their group judgments *relative* to those of other groups. The resulting fractile ratings demonstrate another form of extreme overconfidence: 98% of the subjects believed that their group judgments were in the top half of all group judgments with respect to accuracy. This is in itself an interesting form of overconfidence. One possible explanation is that judges, lacking information about their peer groups, use their own pregrouping level of performance as the standard for other groups. As can be seen by the increased confidence for their group judgment, they had reasons to believe that their group was performing better than this standard. Or, it is possible that high confidence in relative accuracy follows from the high confidence in absolute accuracy seen in the confidence interval data (or vice versa). Finally, it may be that overconfidence about one's group's performance relative to other groups may be related to the use of large monetary incentives for the top 50% of group judgments. Subjects could be demonstrating wishful thinking, attempting to influence the distribution of rewards, or trying to reduce dissonance from voluntary participation (Festinger, 1954).

Of course, changes in accuracy and confidence such as those reported here may be observed for individuals under particular conditions, like successive repetitions of the judgment task. Our goal here is not to identify effects unique to groups, but rather to understand behavior in the situation in which individuals bring their own judgments to the group for

the task of making a single consensus judgment. Finally, it must be cautioned that judgment tasks with ongoing groups may not achieve the same increase in accuracy. The groups used in this study had minimum incentives to strategically distort their assessments to, for example, increase their own influence on the final aggregate group judgment. The incentive scheme for all members of our ad hoc groups encouraged maximum individual and group judgment accuracy. That is, by controlling for effects of history, status, etc., and providing large rewards for accuracy, we hoped to buy honesty in order to obtain participants' maximum performance. The data demonstrate that significant increases in accuracy and confidence in risk judgment can occur through grouping.

REFERENCES

Castore, C. H., & Murnighan, J. K. (1978). Determinants of support for group decisions. *Organizational Behavior and Human Performance*, **22**, 75–92.

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

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

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influence on individual judgment. *Journal of Abnormal and Social Psychology*, **51**, 629–636.

Douglas, M., & Wildavsky, A. (1982). *Risk and culture*. Los Angeles, CA: University of California Press.

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. (1977). Confidence in judgment: Persistence in the illusion of validity. *Psychological Review*, **85**, 395–416.

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

Ferrell, W. R. (1985). Combining individual judgments. In A. Wright (Ed.), *Behavioral decision making*. New York: Plenum.

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

Fischhoff, B., & MacGregor, (1982). Judged lethality: How much people seem to know depends upon how they are asked. *Risk Analysis*, **13**, No. 4, 229–236.

Graesser, C. C. (1982). A social averaging theorem for group decision making. In N. H. Anderson (Ed.), *Contributions to information integration theory*. New York: Academic Press.

Hammond, K. (1980). *The integration of research in judgment and decision theory*. Center for Research on Judgment and Policy, University of Colorado at Boulder, Boulder, CO.

Hammond, K. (1984). Improving scientists' judgments of risk. *Risk Analysis*, **4**(1), 69–78.

Hart, S. L. (1985). Toward quality criteria for collective judgments. *Organizational Behavior and Human Decision Processes*, **36**, 209–228.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Decision research* (Vol. 2). 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.

Janis, I. (1972). *Victims of groupthink*. Boston: Houghton Mifflin.

Kaplan, M. F., & Miller, C. E. (1983). Group discussion and judgment. In P. B. Paulus (Ed.), *Basic group processes*. New York: Springer-Verlag.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, **6**, No. 2, 107–118.

Larson, J. R., Jr., & Reenan, A. M. (1979). The equivalence interval as a measure of uncertainty. *Organizational Behavior and Human Performance*, **23**, 49–55.

Libby, R., Trotman, K. T., & Zimmer, I. Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology*, **72**, No. 1, 81–87.

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

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, **4**, 551–578.

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.

Maier, N. R. F. (1967). Assets and liabilities in group problem solving: The need for an integrative function. *Psychological Review*, **74**(4), 239–249.

Paese, P. W. (1988). Confidence and accuracy in concurrent and predictive judgments of performance. Unpublished doctoral dissertation. University of Illinois at Urbana–Champaign.

Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance*, **24**, 73–92.

Sackman, H. (1974). *Delphi critique*. Lexington, MA: Lexington Books.

Seaver, D. A., von Winterfeldt, D., & Edwards, W. (1978). Eliciting subjective probability distributions on continuous variables. *Organizational Behavior and Human Performance*, **21**, 379–391.

Shanteau, J., & Nagy, G. F. (1982). Information integration in person perception: Theory and application. In M. Cook (Ed.), *Progress in person perception*. London: Metheun.

Sniezek, J. A. (in press). Judgment, confidence, and group process in sales forecasting.

Sniezek, J. A., & Henry, R. A. (in press). Revision, weighting, and commitment in consensus group judgment.

Sniezek, J. A., & Reeves, A. P. (1986). Feature cues in probability learning: Data base information and judgment. *Organizational Behavior and Human Decision Processes*, **37**, 297–315.

Stasser, G., & Davis, J. H. (1981). Group decision making and social influence: A social interaction sequence model. *Psychological Review*, **88**, 523–551.

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

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

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the rationality of choice. *Science*, **211**, 453–458.

Uecker, W. C. (1982). The quality of group performance in simplified information evaluation. *Journal of Accounting Research*, **20**, 388–402.

Winkler, R. L. (1967). The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association*, **62**, No. 320, 1105–1120.

Winkler, R. L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, **66**, No. 336, 675–685.