Journal of Applied Psychology
1995, Vol. 80, No. 4, 443–454

A Comparison of Strategies for Reducing Interval Overconfidence in Group Judgments

S. Plous
Wesleyan University

The present experiments examined several strategies designed to reduce interval overconfidence in group judgments. Results consistently indicated that 3–4-person nominal groups (whose members made independent judgments and later combined the highest and lowest of these estimates into a single confidence interval) were better calibrated than individual judges and interactive groups. This pattern held even when participants were directly instructed to expand their interval estimates, or when interactive groups appointed a devil's advocate or explicitly considered reasons why their interval estimates might be too narrow. Interactive groups did not perform substantially better than individuals, although participants frequently had the impression that group judgments were far superior to individual judgments. This misperception resembles the "illusion of group effectivity" found in brainstorming research.

The relation between judgment accuracy and confidence levels has received increasing attention in applied settings (e.g., Perfect, Watson, & Wagstaff, 1993; Smith, Kassin, & Ellsworth, 1989). It is well documented that people tend to be overconfident of their answers to a wide variety of general knowledge questions—particularly when the questions are difficult (cf. Plous, 1993). For example, Lichtenstein and Fischhoff (1977) found that people were 15%–20% overconfident when the accuracy of their answers was not much better than what would be expected by chance alone. Although early critics claimed that these results were largely a function of asking people about obscure or trivial topics, subsequent studies replicated Lichtenstein and Fischhoff's findings with more commonplace judgments (Dunning, Griffin, Milojkovic, & Ross, 1990; Vallone, Griffin, Lin, & Ross, 1990).

Most studies on overconfidence use one of two approaches. In the first approach, participants are asked to estimate the chances (probability or odds) that their judgments are correct. These estimates are then used to assess the calibration between confidence and accuracy. A person is perfectly calibrated when, across all judgments at a given level of confidence, the proportion of accurate judgments is identical to the expected probability of being correct. In the second approach to studying overconfidence—the approach used in the present report—participants are asked to create "confidence intervals" that have a specific probability (usually .90 or .98) of containing an unknown quantity. The common finding with this approach is that confidence intervals seldom capture the correct answer as often as they should (e.g., 90% confidence intervals contain the correct answer less than 90% of the time). For example, Lichtenstein, Fischhoff, and Phillips (1982) examined several studies in which participants had been asked to give 98% confidence intervals; averaging across all experiments for which information was available—a total of nearly 15,000 judgments—the intervals captured the correct answer only 68% of the time. That is, when participants were 98% sure that an interval contained the correct answer, they were right 68% of the time.

Although findings such as these are typically interpreted as evidence that participants are overconfident, this kind of "overconfidence" may be due primarily to anchoring and insufficient adjustment. Tversky and Kahneman (1974) put it the following way:

> To select X_{90} for the value of the Dow-Jones average, for example, it is natural to begin by thinking about one's best estimate of the Dow-Jones and adjust this value upward. If this adjustment—like most others—is insufficient, then X_{90} will not be sufficiently extreme. A similar anchoring effect will occur in the selection of X_{10}, which is presumably obtained by adjusting one's best estimate downward. Consequently, the confidence interval between X_{90} and X_{10} will be too narrow. (p. 1129)

Thus, unlike the first type of overconfidence—in which

I wish to thank Matthew W. Badger, Irene Kim, Keith Fleming, and Melissa Kim Byun for their assistance in conducting the experiments summarized in this report, as well as H. Diane Ersepke, Keith Fleming, Melissa Kim Byun, and Karl E. Scheibe for their comments on a draft of this article.

Correspondence concerning this article should be addressed to S. Plous, Department of Psychology, Wesleyan University, 207 High Street, Middletown, Connecticut 06459-0408.

participants overestimate the chances they are correct—it is possible for participants to give overly narrow confidence intervals without feeling particularly confident of their answers. Indeed, Sniezek and Buckley (1991) found that many participants who are asked to form 90% confidence intervals later judge their hit rate to be considerably lower than 90%. For this reason, the term *overconfidence* may not accurately describe participants' subjective experience in the case of interval estimation tasks. Nonetheless, following conventional usage, I will refer to the first type of overconfidence as *accuracy overconfidence* and the second type of overconfidence as *interval overconfidence*.

Several studies have explored techniques for reducing accuracy overconfidence. In one pair of experiments, Lichtenstein and Fischhoff (1980) found that people who were initially overconfident could learn to be better calibrated after making 200 judgments and receiving intensive performance feedback. Likewise, Arkes, Christensen, Lai, and Blumer (1987) found that overconfidence could be eliminated by giving participants feedback after five deceptively difficult problems. These studies show that overconfidence can be unlearned, although their applied value is somewhat limited. Few people will ever undergo training sessions to become well calibrated.

A more practical debiasing strategy was tested by Koriat, Lichtenstein, and Fischhoff (1980). In this research, participants answered two sets of two-alternative general knowledge questions, first under *control* instructions and then under *reasons* instructions. Under control instructions, participants chose an answer and estimated the probability (between .50 and 1.00) that their answer was correct. Under reasons instructions, they were asked to list reasons for and against each of the alternatives before choosing an answer. Koriat et al. found that participants were overconfident when given control instructions but became very well calibrated after generating pro and con reasons (comparable to participants who were given intensive feedback in the study by Lichtenstein and Fischhoff, 1980). Also, in a follow-up experiment, Koriat et al. found that the generation of opposing reasons alone was sufficient to reduce accuracy overconfidence. Unfortunately, however, subsequent research on this technique has yielded mixed results (Fischhoff & MacGregor, 1982; Hoch, 1985; Trafimow & Sniezek, 1994).

Moreover, no comparable technique has been tested for reducing interval overconfidence. To date, only two published studies have assessed debiasing strategies for reducing interval overconfidence. The first study, conducted by Alpert and Raiffa (1982), examined the effectiveness of performance feedback. After participants showed typical levels of overconfidence in response to 10 sociometric and general knowledge questions (e.g., number of eggs produced annually in the United States), they were given detailed performance feedback urging them to "Be honest with yourselves! Admit what you don't know! . . . Spread out those distributions!" Participants then answered a second set of 10 questions, but despite being prodded, they showed only a modest improvement in calibration. The second study, conducted by Block and Harper (1991), was only peripherally concerned with methods for reducing overconfidence. In one experiment in this series, half of the participants were explicitly warned "not to be overly confident in your knowledge." Once again, however, warnings had relatively little effect on confidence levels.

These results attest to the persistence of interval overconfidence in individual judgments, yet they say little about interval overconfidence in group judgments. Do groups show the same degree of interval overconfidence as individuals? Are warnings about interval overconfidence as ineffectual with groups as with individuals? And if so, are there other measures that can be taken to improve the calibration of group judgments? Such questions are especially important in light of the frequency with which groups are called on to provide interval estimates. For example, corporate planning groups routinely use interval estimates in sales forecasts, profit projections, market appraisals, and so forth. Likewise, military advisory panels frequently use interval estimates when assessing troop strength, weapons stockpiles, and a variety of other uncertain quantities. Miscalibration in these cases can lead to financial and military disasters.

Despite the prevalence of interval estimates in applied settings, however, relatively few calibration studies have directly compared confidence in group judgments with confidence in individual judgments (Sniezek, 1992). This omission is particularly curious given the voluminous research that has been conducted on groupthink, risky shift phenomena, and group polarization. Furthermore, several studies have found that groups tend to be more confident than individuals, and that group discussion tends to heighten previous confidence levels (Ono & Davis, 1988; Seaver, 1979; Sniezek & Henry, 1989; Tindale, 1983). For example, Stephenson and his colleagues found that groups are more prone than individuals to a misplaced confidence in inaccurately recalled material (Clark, Stephenson, & Kniveton, 1990; Stephenson, Brandstätter, & Wagner, 1983; Stephenson, Clark, & Wade, 1986). Results such as these suggest the value of studying overconfidence at the group level.

In the present experiments, I explored the effectiveness of several techniques designed to reduce interval overconfidence in 3–4-person groups. In Experiment 1, individual judgments were compared with judgments based on group consensus and judgments formed by pooling the highest and lowest estimates of group members working separately. Experiment 2 replicated the findings of Experiment 1 with survey items higher in familiarity and personal relevance than those used in Experiment 1. Experiment 3 compared the effectiveness of three different techniques: (a) the pooling method used in Experiment 1, (b) a technique based on the devil's advocacy approach to group decision making, and (c) a technique involving the consideration of reasons why the group's answers might be wrong. Experiment 4 extended previous results (based on 90% confidence intervals) to the case of 75% confidence intervals, and examined the goals and strategies used by participants.

Experiment 1

Research on group problem solving has often shown that nominal, or "statisticized," groups (whose members work alone and later pool their ideas) outperform similarly sized groups of interacting individuals (Diehl & Stroebe, 1987; Guzzo, 1986; Hill, 1982). According to the results of one recent meta-analysis, nominal groups exceed interactive groups by an average of roughly two thirds of one standard deviation unit on both quantitative and qualitative indices of performance (Mullen, Johnson, & Salas, 1991). These results raise the question of whether a similar pattern occurs with respect to interval estimation. Specifically, the question is this: If a small number of people are faced with the task of providing well-calibrated confidence intervals, would they be better off working alone (and later pooling their independent estimates) or working together as a group?

Some judgment biases are attenuated through group discussion (E. F. Wright, Lüüs, & Christie, 1990; E. F. Wright & Wells, 1985), and others are not (Argote, Seabright, & Dyer, 1986; Tindale, Sheffey, & Filkins, 1990). Although groups usually have more information than individuals and are often able to catch mistakes (Hill, 1982), the process of group discussion does not always lead to an efficient sharing of information (Stasser & Titus, 1985, 1987), and in some instances leads to a polarization or amplification of individual-level biases (Argote et al., 1986; Myers & Lamm, 1976). If interval overconfidence arises mainly from anchoring and insufficient adjustment, nominal groups should be better calibrated than interactive groups. By having each member generate estimates separately, nominal groups avoid the pitfall of having one group member suggest an answer that serves as an anchor during group discussions. In addition, nominal groups have the advantage of avoiding conformity pressures that inhibit the expression of dissenting viewpoints.

There are many different ways to pool interval estimates within nominal groups, including various averaging schemes, the expansion of intervals by a fixed percentage, and so forth. After extensive pilot testing, the method adopted in the present set of experiments was to form confidence intervals from the highest and lowest individual estimates given for a particular item. This method produces larger confidence intervals than methods that use average upper and lower bounds, and it avoids certain logistical problems in expanding intervals by a fixed percentage (e.g., if a participant were to estimate Mozart's year of death somewhere between 1650 and 1850, mechanically doubling the width of this interval around a midpoint of 1750 would lead to a nonsensical upper estimate of 1950; similarly, if a participant were to estimate the coastline of New Hampshire at 0–20 miles, expanding the interval width by 50% would lead to a meaningless lower estimate of −5 miles). The main objective was to see whether pooling the high and low estimates of a few independent judges would be sufficient to correct for individual-level overconfidence and idiosyncratic blind spots in knowledge, thereby providing an easy and economical means for achieving well-calibrated confidence intervals.

I designed Experiment 1 to see whether this pooling strategy produces judgments that are better calibrated than interactive group judgments, and to see whether participants are sensitive to differences in performance among individual, nominal, and interactive modes of estimation. In a repeated measures design, consensus judgments on one survey were compared with pooled individual judgments on a comparable survey. Judgments were predicted to be better calibrated when the group pooled independent estimates than when it arrived at consensus estimates through discussion. In a posttask questionnaire, participants were also asked to guess how many items they answered correctly when working alone, when pooling their answers, and when discussing items together as a group. The purpose of the posttask questionnaire was to see whether participants were aware of differences in their performance under various conditions.

Method

Participants. Eighty-five college students took part in Experiment 1. Seventy-seven students participated to fulfill an introductory psychology course requirement, and 8 students were paid volunteers recruited through a school newspaper advertisement. In all, 53 women and 32 men participated.

Judgment task. Two surveys (Survey A and Survey B) were developed for use in Experiment 1. Survey A asked participants to provide 90% confidence intervals for each of 10 almanac items taken from Russo and Schoemaker (1989). Participants were instructed to provide a low and high guess for each item such that they were 90% sure the correct answer fell between the two, and they were warned: "When most people attempt this task, they give ranges that are far too narrow. Don't fall into the same trap!" The directions explained that participants' ranges should be "neither too narrow nor too wide," that perfect calibration would mean missing exactly 10% of the items (i.e., one question), and that participants would be entered into a $10 cash prize lottery if they got exactly 9 of the 10 items correct.

Survey B served as an alternate form of Survey A (to allow for within-subject comparisons). Survey B was identical to Survey A in all respects except the subject matter of its almanac items. In constructing Survey B, primary consideration was given to the difficulty of the questions and the magnitude of the answers (each answer in Survey B was within 10% of a corresponding item in Survey A). Survey B was pretested on a separate group of 45 participants, and was found to be roughly equivalent in difficulty to Survey A. A comparison of survey items can be found in Table 1.

Procedure. Twenty-five experimental sessions were conducted, each with a 3–4-person group that had been randomly assigned to one of four conditions (roughly the same number of 3-person and 4-person groups were used in this experiment and those that follow, and no systematic differences were found based on group size; hence, this factor is not discussed further).

Table 1
Experiments 1, 3, and 4: Survey Items and Correct Answers

Item	Correct answer
1A. Diameter of the moon (in miles)	2,160
1B. Length of the Great Wall of China (in miles)	2,150
2A. Year in which Wolfgang Amadeus Mozart was born	1756
2B. Year in which Sir Isaac Newton died	1727
3A. Number of countries that are members of OPEC	13
3B. Miles of ocean coastline in New Hampshire	13
4A. Air distance from London to Tokyo (in miles)	5,959
4B. Air distance from Los Angeles to Rio de Janeiro, Brazil (in miles)	6,330
5A. Number of books in the Old Testament	39
5B. Number of existing hazardous waste sites in Minnesota (according to the EPA)	41
6A. Martin Luther King's age at death	39
6B. Gestation period (in days) of a kangaroo	42
7A. Gestation period (in days) of an Asian elephant	645
7B. Number of earth days in a Mars year	687
8A. Length of the Nile River (in miles)	4,167
8B. Number of accidental deaths per year in the U.S. from fires and burns	4,400
9A. Weight of an empty Boeing 747 (in pounds)	390,000
9B. Combined population of Bridgeport, Hartford, and New Haven, Connecticut	407,027
10A. Deepest (known) point in the oceans (in feet)	36,198
10B. Number of farms in New York state, according to the United States Department of Agriculture	39,000

Note. Items marked *A* appeared in Survey A and were taken from *Decision traps: Ten barriers to brilliant decision making and how to overcome them* (p. 71), by J. E. Russo and P. J. H. Schoemaker. Copyright © 1989 by J. Edward Russo and P. J. H. Schoemaker. Used by permission of Doubleday, a division of Bantam Doubleday Dell Publishing Group, Inc. Items marked *B* appeared in Survey B (developed for the present study).

In two conditions participants completed Survey A followed by Survey B, and in two conditions they completed Survey B followed by Survey A. Crossed with this factor, approximately half of the participants first worked individually and later worked in groups, and approximately half worked in groups and later worked individually. When participants worked alone, they were given 5–10 min to complete one of the surveys (A or B, depending on the condition). Then, after participants finished their surveys, they were seated together and were asked to combine their estimates for each item. Participants were instructed, "For each of the following ten items, please fill in the lowest value anyone in your group gave for the LOW estimate, and the highest value anyone gave for the HIGH estimate."

When participants worked interactively in groups, they were seated together and instructed to arrive at a consensus for each range estimate. Interactive groups were given approximately 10 min to complete one of the two survey forms (again, A or B, depending on the condition). As a performance incentive, participants in each group were told that if their group got nine items correct on the survey, each member would be entered in a drawing for $10. Thus, the incentive ($10) and time allotment (10 min) were roughly the same whether participants worked in a group or worked separately and later pooled their answers (thereby eliminating these factors as potential confounds in subsequent performance comparisons).

In all, then, the experiment had the following four conditions: (a) individual: Survey A, group: Survey B (seven groups); (b) individual: Survey B, group: Survey A (six groups); (c) group: Survey A, individual: Survey B (six groups); and (d) group: Survey B, individual: Survey A (six groups). Once participants had completed individual surveys (pooling their independent estimates) and a group survey, they were asked to estimate the number of items they got correct on (a) the survey they completed alone, (b) the survey form that combined low and high estimates for the group, and (c) the survey they completed as a group. The inclusion of these posttask estimates was based on the recommendations of Sniezek and Buckley (1991), who found that participants frequently use different processes to generate confidence intervals and posttask performance appraisals, and who suggested that both measures be used when assessing confidence levels.

Results and Discussion

Pooling versus interactive discussion. In Experiment 1, groups were used as the main unit of analysis. A repeated measures analysis of variance (ANOVA) on group survey scores was conducted with two between-groups variables (whether participants first worked individually or first worked in a group, and whether they were first given Survey A or Survey B) and one within-subject variable (whether the scores were arrived at by pooling individual estimates or by discussing the items in a group). This analysis showed a highly significant effect on the within-subject variable, $F(1, 21) = 157.15, p < .001$, and no other significant effects or interactions (i.e., there were no significant effects or interactions resulting from survey form or task ordering). When participants worked together, they averaged 4.1 items correct, a significant but limited improvement over the average score achieved by participants working alone (3.2 items), $t(24) = 2.85$, $p < .009$. These results agree closely with those found by Russo and Schoemaker (1992), who asked 83 managers to provide 90% confidence intervals for 10 unknown quantities and then discuss their answers in groups of 3 or 4 participants. Managers working alone averaged 2.8 items correct; working in groups, they averaged 4.4 items correct.

On the other hand, when participants in the present study pooled their low and high estimates, they achieved an average score of 7.7 items correct—nearly twice the average of groups working interactively (though even then, scores fell significantly below the perfect calibration mark of 9 items correct, $t(24) = 4.80$, $p < .001$). Looking at the data another way, the pooling strategy yielded a score of 7–9 correct items 84% of the time, whereas group interactions led to such scores only 8% of the time. These results suggest that pooling provides a simple method for improving the calibration of interval estimates, whereas interactive group judgments are vulnerable to the same kind of anchoring and overconfidence observed at the individual level. As 1 participant remarked in a postexperimental interview, "Usually someone would throw something out and we'd work with it from there."

Posttask performance estimates. Following completion of the surveys, participants were asked to estimate the number of items they had gotten correct when (a) working alone, (b) working in a group, and (c) combining the separate estimates of people in their group. As shown in Figure 1A, the first two of these estimates differed markedly from participants' true scores. On the average, participants guessed that they had obtained a score of 5.2 while working alone, when in fact they averaged 3.2, $t(84) = 9.47$, $p < .001$, and they estimated their group score at an average of 7.0, when groups actually averaged 4.1, $t(24) = 9.08$, $p < .001$. In both of these cases participants overestimated their accuracy, though they did not go so far as to report having gotten 9.0 items correct. These figures are similar to those reported by Sniezek and Buckley (1991), who found that on the average, individual participants forming ten 90% confidence intervals guessed that they had gotten 5.6 items correct. In contrast to this pattern, participants were extremely accurate when estimating the number of items correct after pooling their answers. The mean estimate was 8.0 items correct, nearly identical to the actual average of 7.7, $t(24) = 1.02$, *ns*. Thus, participants were not only better calibrated when they used the pooling strategy, but their posttask performance appraisals were more accurate as well.

Experiment 2

The pooling strategy tested in Experiment 1 yielded far better calibration than did group discussion. When groups of participants pooled their most extreme range estimates, they averaged 7.7 items correct out of 10—nearly double the number of items interactive groups got correct. These results suggest that pooling three or four independent confidence intervals by combining the highest and lowest individual range estimates may constitute an effective method for reducing interval overconfidence.

Such a conclusion is limited, however, by the type of questions used in Experiment 1. Most questions involved topics with which participants had little familiarity (e.g., air distance from London to Tokyo). Because several authors have questioned the generality of results based exclusively on general knowledge questions (e.g., Ronis & Yates, 1987), and others have found that overconfidence diminishes when participants are asked about a familiar domain (Block & Harper, 1991; Russo & Schoemaker, 1992), I conducted a second experiment using questions of greater immediacy to a student population. In this experiment, students were asked to provide range estimates for a variety of unknown quantities concerning their university.

Method

Participants. The participants were 97 college students at a small liberal arts institution (Wesleyan University). Fifty-seven students participated to fulfill an introductory psychology course requirement, and 40 students were paid volunteers recruited through a school newspaper advertisement. In all, 49 women and 48 men participated.

Procedure. Experiment 2 used the same design and procedure as Experiment 1. The only notable difference was that the surveys used in Experiment 1 were replaced with two new surveys. These surveys asked students to provide 90% confidence intervals for various attributes of their university (e.g., number of students living off campus, number of tenured Latino professors, and so on). As before, Form A and Form B contained items matched in magnitude, and pilot tests indicated that both forms were comparable in difficulty. Twenty-nine groups were randomly assigned as follows: (a) individual: Survey A, group: Survey B (eight groups); (b) individual: Survey B, group: Survey A (seven groups); (c) group: Survey A, individual: Survey B (eight groups); and (d) group: Survey B, individual: Survey A (six groups).

Results and Discussion

As in Experiment 1, groups were treated as the main unit of analysis, and as before, a repeated measures ANOVA showed that the pooling strategy yielded higher scores than group interaction, $F(1, 25) = 102.13$, $p < .001$. When participants pooled their low and high estimates, they averaged 6.8 items correct, compared with

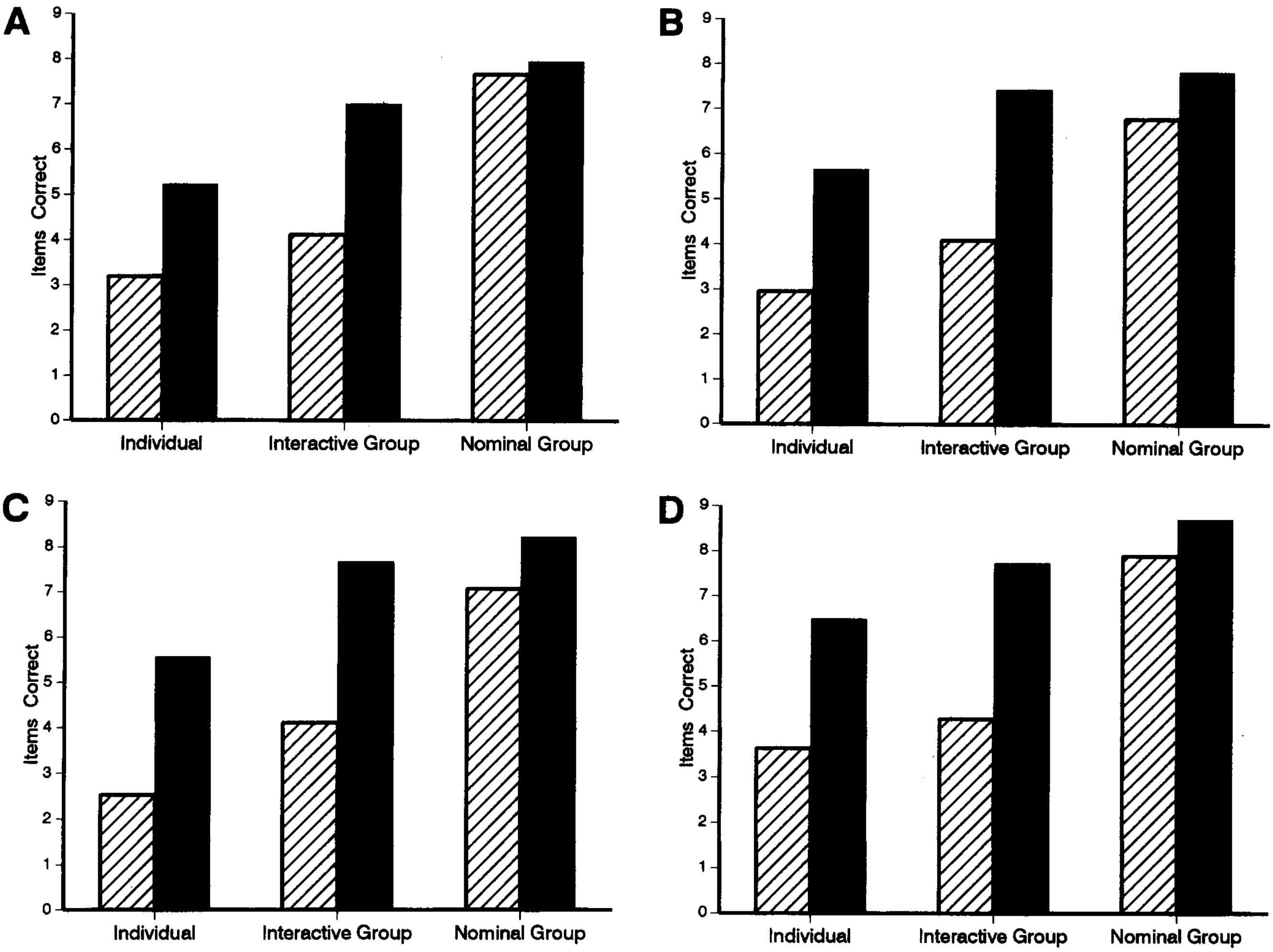

Figure 1. A comparison of the mean number of items correct (striped bars) and posttask performance estimates (solid bars) for (A) Experiment 1, (B) Experiment 2, (C) Experiment 3, and (D) Experiment 4.

4.1 correct when working interactively in groups and 3.0 when working alone. Although 6.8 is almost 1 item less than the figure of 7.7 found in Experiment 1, the other two averages are virtually identical to those found in Experiment 1 (4.1 and 3.2, respectively), and the overall pattern of results in Experiment 2 was the same as that found in Experiment 1 (see Figure 1B). In general, then, Experiment 2 replicated the earlier findings from Experiment 1.

Experiment 2 also replicated the same basic pattern of posttask estimates found in Experiment 1. On the average, participants guessed they had gotten 7.8 items correct when pooling their range estimates, fairly close to their actual average of 6.8 (though unlike Experiment 1, this difference reached statistical significance, $t(28) = 2.99$, $p < .006$). In contrast, participants substantially overestimated how well they did when performing alone (5.7 vs. 3.0), $t(96) = 13.91$, $p < .001$, and in interacting groups, (7.4 vs. 4.1), $t(28) = 11.32$, $p < .001$. Thus, as in Experiment 1, participants were not only better calibrated when using the pooling strategy, but they were also more accurate in appraising their performance.

Experiment 3

The results of Experiments 1 and 2 suggest that pooled interval estimates are often better calibrated than judgments arrived at through group discussion. In both experiments, nominal groups were relatively well calibrated and were quite accurate in appraising their performance, whereas interacting groups averaged only 4.1 items correct and later guessed that they had performed much better. Before concluding that group interaction is not productive, however, it is worth considering the possibility that alternative forms of group interaction might lead to improved calibration. I designed Experiment 3 to test this possibility. In this experiment, participants began by working alone and pooling their estimates. Then they were given a second survey and asked to work as a group in one of two ways: (a) by appointing a devil's advocate to challenge the group's answers or (b) by explicitly considering reasons why the group's answers might be wrong. These group discussion techniques are based on well-known debiasing strategies used in previous research (e.g., Koriat et al., 1980; Schwenk & Cosier, 1980).

Method

Participants. Ninety-two college students took part in Experiment 3. Fifty-two students participated to fulfill an introductory psychology course requirement, and 40 students were paid volunteers recruited through a school newspaper advertisement. In all, 57 women and 35 men participated.

Procedure. Groups of 3–4 participants answered the same survey items used in Experiment 1. After initially completing either Survey A or Survey B alone, participants were assembled as a group and asked to combine their highest and lowest estimates for each item. Next, participants were given 15 min to complete the alternate survey as a group (Survey A if they had been given Survey B initially, or Survey B if they had been given Survey A). Participants in the *devil's advocacy* condition were told, "When most people attempt this task, they give ranges that are far too narrow. Don't fall into the same trap! To protect against overconfidence, choose one member of your group to play the role of 'devil's advocate.' This person should get the group to consider reasons why its range estimates might be too narrow." Participants in the *reasons* condition were given a modified version of the debiasing instructions used by Koriat et al. (1980), "When most people attempt this task, they give ranges that are far too narrow. Don't fall into the same trap! Instead, consider reasons why your range might be too narrow. Such reasons might include facts that you know, things that you vaguely remember, assumptions that make you believe the range is too small, gut feelings, associations and the like." In all, 27 groups were randomly assigned as follows: (a) individual: Survey A, devil's advocacy: Survey B (seven groups); (b) individual: Survey B, devil's advocacy: Survey A (seven groups); (c) individual: Survey A, reasons: Survey B (seven groups); and (c) individual: Survey B, reasons: Survey A (six groups). As in Experiments 1 and 2, participants were offered entry into a $10 cash lottery if they got exactly nine items correct, and they were asked after the experiment to estimate how many items they had gotten correct when working alone, pooling their estimates, and working as a group.

Results and Discussion

A repeated measures ANOVA on group survey scores was conducted with two between-groups variables (whether participants answered Survey A first or Survey B first, and whether their group was instructed to appoint a devil's advocate or to consider reasons why its answers might be wrong) and one within-subject variable (whether scores were arrived at through pooling or group discussion). This analysis showed a highly significant effect on the within-subject variable, $F(1, 23) = 63.32, p < .001$, and no other significant effects or interactions. As in Experiments 1 and 2, participants working together averaged precisely 4.1 items correct—a small but significant improvement over individual performance ($M = 2.5$), $t(26) = 3.92, p < .001$, though well below the average of 7.1 achieved by pooling individual estimates, $t(26) = 7.97, p < .001$. Thus, the debiasing instructions tested in Experiment 3 did not improve group performance beyond the level observed in Experiment 1, and the superiority of the pooling strategy was strongly replicated.

As seen in Figure 1C, the posttask estimates also followed the same pattern observed previously. On average, participants guessed they had gotten 8.2 items correct when pooling their range estimates, fairly close to their actual average of 7.1 (though this difference was significant, $t(26) = 4.33, p < .001$). In contrast, participants working alone or in interactive groups thought they had gotten roughly twice as many items correct as they actually had. On average, participants working alone guessed that they had gotten 5.6 items correct when they had actually gotten 2.5 correct, $t(91) = 14.80, p < .001$, and participants working interactively guessed that they had gotten 7.8 items correct when they had actually averaged 4.1, $t(26) = 9.33, p < .001$. As before, these findings suggest that participants not only achieve better calibration when using the pooling strategy, but that their posttask performance appraisals are more accurate.

Experiment 4

Although the results of Experiment 3 strongly replicate Experiments 1 and 2, they are limited in two respects. First, participants were allotted a maximum of 15 min for group discussion, which may not have been long enough for some groups to fully realize the benefits of a debiasing strategy such as devil's advocacy. Second, the findings of Experiment 3 (as well as Experiments 1 and 2) are limited exclusively to 90% confidence intervals, yet extreme confidence intervals such as this may produce atypical levels of overconfidence and may exaggerate the usefulness of mechanical pooling strategies (Roth, 1993; Solomon, 1982). The purpose of Experiment 4 was to remedy these limitations, and to further explore the goals, strategies, and thoughts participants had while making interval estimates.

Method

Participants. One hundred college students took part in Experiment 4. Forty-five students participated to fulfill an introductory psychology course requirement, and 55 students were paid volunteers. In all, 58 women and 42 men participated.

Procedure. Twenty-eight groups of 3–4 participants answered the same survey items used in Experiments 1 and 3. All participants first answered Survey A or B together using the devil's advocacy approach outlined in Experiment 3, then completed the alternate survey individually, and then integrated their individual estimates as in previous experiments. The survey instructions and performance incentives were similar to those used in Experiment 3, except that participants were explicitly given an unlimited amount of time to discuss their estimates when using the devil's advocacy approach. Also, half of the participants were asked to form 90% confidence intervals (both when working as a group and when working individually), and the other half were asked to form 75% confidence intervals (i.e., to get 7 or 8 items correct out of 10). This variable was crossed with the order of survey presentation, yielding four different experimental conditions: (a) Survey A first, 90% intervals (seven groups); (b) Survey B first, 90% intervals (seven groups); (c) Survey A first, 75% intervals (seven groups); and (d) Survey B first, 75% intervals (seven groups).

At the conclusion of the experiment, participants were given a posttask questionnaire in which they estimated how many items they had gotten correct when working alone, pooling their estimates, and working as a group. The questionnaire also asked participants to identify what their goal had been when working alone, what their strategy had been when working alone, what their goal had been when working in a group, and what their strategy had been when working in a group. The questions on goals offered four alternatives: (a) "To get all the questions correct," (b) "To get exactly 9 questions correct," (c) "To get 7 or 8 questions correct," or (d) "Other—please explain." The questions on strategy offered the following alternatives: (a) "Try to get as many items correct as possible," (b) "Try to get certain items correct and certain item(s) wrong," (c) "Try to feel 90% [75%] sure that each item is correct," or (d) "Other—please explain." Finally, in a free-response question, participants who estimated that they had gotten fewer items correct than the number specified as their goal (either individually or as a group) were asked to explain why they had not widened their interval estimates to get more items correct.

Results and Discussion

On average, participants took 19.8 min ($SD = 5.44$) when offered an unlimited amount of time for group discussion. To analyze group survey scores, I conducted a repeated measures ANOVA with two between-groups variables (whether participants answered Survey A first or Survey B first, and whether they gave 90% confidence estimates or 75% confidence estimates) and one within-subject variable (whether scores were arrived at through pooling or group discussion). This analysis showed a highly significant effect on the within-subject variable, $F(1, 24) = 89.74, p < .001$, and no significant main effect for order of survey presentation or confidence interval width (there was, however, a significant Order × Interval Width interaction, in which 90% confidence estimates were correct more often than 75% confidence estimates when Survey B was given first, but 75% confidence estimates were correct more often than 90% confidence estimates when Survey A was given first, $F(1, 24) = 6.33, p < .02$). Participants averaged 4.3 items correct when using a devil's advocacy approach—only marginally better than when working alone ($M = 3.6$), $t(27) = 1.83, p < .08$. In contrast, when pooling their estimates, participants averaged 7.9 items correct. More specifically, participants giving 75% confidence estimates averaged 7.8 items correct after pooling, which was not significantly different than the perfect calibration mark of 7.5 items, $t(27) = 0.75$, *ns*, and participants giving 90% confidence intervals averaged 8.0 items correct, a small but significant difference from the perfect calibration mark of 9 items correct, $t(27) = 3.18, p < 008$. These results replicate the findings of Experiments 1–3, and they extend previous findings to the case of 75% confidence estimates.

As shown in Figure 1D, posttask performance appraisals also followed the same pattern observed previously. On average, participants guessed they had gotten 8.7 items correct when pooling their range estimates, a small but significant overestimate of their actual 7.9 item average, $t(27) = 2.50, p < .02$. In contrast, participants working alone or in interactive groups thought they had gotten nearly twice as many items correct as they actually had. On average, participants working alone guessed that they had gotten 6.5 items correct when they had actually gotten 3.6 correct, $t(99) = 13.18, p < .001$, and participants working together in a group guessed that they had gotten 7.7 items correct when they had actually averaged 4.3, $t(27) = 8.34, p < .001$. Thus, as before, pooling not only led to better calibration but to more accurate posttask performance appraisals.

Perhaps the most surprising finding of Experiment 4 is that participants in the 90% conditions did not get significantly more items correct than participants in the 75% conditions. One explanation for this finding may lie in the goals and strategies participants used in the estimation task. As seen in Table 2, even though 90% participants and 75% participants differed in their goals, $\chi^2(3, N = 100) = 23.26, p < .001$, the most common goal in both cases was to get all 10 questions correct—an objective pursued by 72% of participants (these results, and the data in Table 2, concern only goals and strategies used in group discussion; because the goals and strategies participants used when working alone were highly similar to those used when working in groups, an analysis of this data is omitted to avoid redundancy). Furthermore, 90% participants and 75% participants did not differ significantly in the strategies they used to make their interval estimates. In both instances, the most common strategy participants adopted was to get as many items correct as possible (sometimes as a hedge, with the realization that at least one item would probably be wrong anyway; more

Table 2
Frequency of Reported Goals and Strategies Used by Groups in Experiment 4

Goal and strategy	Condition	
	75% confidence	90% confidence
Goal		
Get all questions correct	35	37
Get 9 questions correct	0	10
Get 7–8 questions correct	13	0
Other goals	2	3
Strategy		
Get as many items correct as possible	28	32
Feel 75% [90%] sure of each item	19	17
Other strategies	3	1

Note. These frequencies are based on the self-reports of 100 participants: 50 who gave 75% confidence intervals, and 50 who gave 90% confidence intervals.

often, simply for the challenge of trying to get 100% of the items correct). Whether working alone or in a group, approximately 60% of all participants reported using this strategy.

These findings suggest that 90% participants and 75% participants achieved the same general level of accuracy because they used the same strategy: namely, the strategy of pursuing 100% accuracy. Although in one sense these results weaken the validity of the experiment (in that a majority of participants were not following the directions precisely), in another sense they render the failure of devil's advocacy even more dramatic; if group members were attempting to achieve 100% accuracy, their interval estimates reflect even greater overconfidence than previously assumed.

Given the fact that many participants strove for 100% accuracy, the question remains as to why participants did not widen their confidence intervals when they suspected the intervals were too narrow. Participants were queried on this topic in the posttask questionnaire, and their answers were both varied and illuminating. Some participants reported a fear that unduly wide ranges would make them look ignorant. Others reported that anyone could give enormously wide intervals, but that the trick was to capture the true value as closely as possible. Some participants even thought that giving wider intervals would be a form of cheating because the instructions had said to give estimates that were "neither too narrow nor too wide." And several participants indicated that extremely large intervals would be meaningless, uninformative, or unrewarding. For example, 1 participant wrote, "I felt like it would be a sort of cop out to just give a huge range that was sure to contain the correct number. . . . I'd rather give a good guess and be wrong with a narrower range than give a huge range and be right." Similar trade-offs between accuracy and informativeness have been observed by Yaniv and Foster (1992).

General Discussion

Of the debiasing methods examined in the present research, only the pooling technique yielded interval estimates that were reasonably well calibrated. Overconfidence persisted in the face of explicit warnings, instructions to expand interval widths, and extended group discussion. Indeed, even when groups were equipped with debiasing procedures such as the devil's advocacy approach, they did not perform substantially better than individuals working alone.

It is interesting to speculate as to why groups of 3 or 4 interacting participants did not do much better than participants working alone. Sniezek (1992) has proposed that groups exhibit less overconfidence than individuals when two conditions are met: (a) The group has more information collectively than its separate members have alone and (b) this information is shared and processed during group interaction. Although the first condition was surely met in the present experiments (cf. Kaplan & Miller, 1983), there is reason to doubt that the second condition was. As alluded to earlier, conformity pressures and anchoring biases were apparent during group discussions, and the estimates offered by one group member may have served to inhibit other group members from sharing their own estimates. This pattern is consistent with the results of Sniezek, Paese, and Furiya (1990), who found that less than a third of all individual judgments were shared during similar small group discussions. Likewise, Stasser and his colleagues have amassed an impressive body of evidence suggesting that group members often fail to exchange unshared information (e.g., Stasser, Taylor, & Hanna, 1989; Stasser & Titus, 1985, 1987). As Stasser and Titus (1987, p. 92) put it, "face-to-face, unstructured discussion while trying to reach a consensus is a poor way for members to inform one another of previously unconsidered information."

The present results are also consistent with brainstorming research that has shown an "illusion of group effectivity" (Diehl & Stroebe, 1991; Paulus, Dzindolet, Poletes, & Camacho, 1993; Stroebe, Diehl, & Abakoumkin, 1992). In these studies, participants believe that their performance is greatly enhanced through group interaction, though such is actually not the case. For example, despite the fact that nominal groups typically outproduce brainstorming groups by a margin of nearly two to one (Paulus et al., 1993), 80% of the participants in one study indicated that a person working in a group would be more effective than a person working alone, and participants tended to feel that the presence of others facilitated their own performance (Stroebe et al., 1992).

In a similar vein, interactive groups in the present experiments displayed almost the same degree of interval overconfidence as lone individuals, yet after rendering their estimates, participants often felt that their group judgments had been far superior to their individual judgments. In other words, when asked after the estimation task to guess how many items they had gotten correct, participants overestimated their group performance by a greater margin than their individual or pooled performance. This pattern, found in all four experiments, suggests that members of a group may feel they have benefited substantially from group discussions, when in reality, relatively little benefit has accrued. Although not specifically tested in the present experiments, this misperception may be the result of an unwarranted assumption that the greater collective knowledge possessed by groups will automatically translate into more accurate answers.

Finally, it is worth considering a few limitations of the present work. First, despite the usefulness of the pooling strategy, it should be noted that nominal groups often fell short of perfect calibration. Even when the separate estimates of four people were pooled, a small degree of overconfidence was frequently evident. In Experiment 1, the mean number of items correct was 8.2, in Experiment 2 the mean was 7.5, in Experiment 3 the mean was 7.2, and in Experiment 4 the mean was 8.1. Thus, when very difficult estimates must be made with high confidence—the type of estimates that, as in the present case, often elicit extreme overconfidence—it may be best to pool more than three or four estimates. Groups of three or four were chosen in the present studies simply because they are common in business and other applied settings, because they are easier to assemble than larger groups, and because the value of additional estimates tends to diminish under pooling strategies (Libby & Blashfield, 1978).

Second, the present findings are limited by the type of questions asked. Although participants in Experiment 2 were asked about a topic they were interested in and knew something about, the questions still took much the same form as almanac questions. As Ronis and Yates (1987) have pointed out, however, overconfidence depends in part on the type of judgments that are being made, and general knowledge questions tend to produce relatively high degrees of overconfidence (see also G. Wright & Wisuda, 1982). Juslin (1994) has even argued that overconfidence on almanac problems is an experimental artifact arising from item selection biases. Future research should therefore examine whether the present results apply to other types of judgments.

Third, only a small number of interactive group techniques were explored: unstructured discussions to consensus, devil's advocacy, and a consider-the-opposite debiasing technique. Yet there may be other interactive techniques that reduce interval overconfidence more effectively than these. Such techniques include the Delphi method, the best member technique, nominal-interactive hybrids such as the nominal group technique devised by Delbecq and Van de Ven, and many others (Gustafson, Shukla, Delbecq, & Walster, 1973; McGrath, 1984; Rohrbaugh, 1979, 1981; Sniezek, 1989; Van de Ven & Delbecq, 1971). Thus, it would be premature to conclude that pooling is superior to all forms of group interaction.

Finally, as noted earlier, calibration is only one index of performance, and it need not be the most important index in every case. In some instances, for example, decision makers may be willing to accept a reduced level of calibration in exchange for more informative (i.e., narrower) interval estimates (Yaniv & Foster, 1992). Consequently, future work on debiasing techniques should consider other performance indices beyond calibration, and should explore methods to improve calibration without unduly sacrificing informativeness.

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). Cambridge, England: Cambridge University Press.

Argote, L., Seabright, M. A., & Dyer, L. (1986). Individual versus group use of base-rate and individuating information. *Organizational Behavior and Human Decision Processes, 38*, 65–75.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133–144.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*, 188–207.

Clark, N. K., Stephenson, G. M., & Kniveton, B. H. (1990). Social remembering: Quantitative aspects of individual and collaborative remembering by police officers and students. *British Journal of Psychology, 81*, 73–94.

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology, 53*, 497–509.

Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of Personality and Social Psychology, 61*, 392–403.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology, 58*, 568–581.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting, 1*, 155–172.

Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Performance, 9,* 280-291.

Guzzo, R. A. (1986). Group decision making and group effectiveness in organizations. In P. S. Goodman (Ed.), *Designing effective work groups* (pp. 34-71). San Francisco: Jossey-Bass.

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

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 719-731.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226-246.

Kaplan, M. F., & Miller, C. E. (1983). Group discussion and judgment. In P. B. Paulus (Ed.), *Basic group processes* (pp. 65-94). New York: Springer-Verlag.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107-118.

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. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159-183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149-171.

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* (pp. 306-334). Cambridge, England: Cambridge University Press.

McGrath, J. E. (1984). *Groups: Interaction and performance.* Englewood Cliffs, NJ: Prentice-Hall.

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology, 12,* 3-23.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602-627.

Ono, K., & Davis, J. H. (1988). Individual judgment and group interaction: A variable perspectives approach. *Organizational Behavior and Human Decision Processes, 41,* 211-232.

Paulus, P. B., Dzindolet, M. T., Poletes, G., & Camacho, L. M. (1993). Perception of performance in group brainstorming: The illusion of group productivity. *Personality and Social Psychology Bulletin, 19,* 78-89.

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology, 78,* 144-147.

Plous, S. (1993). *The psychology of judgment and decision making.* New York: McGraw-Hill.

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

Rohrbaugh, J. (1981). Improving the quality of group judgment: Social judgment analysis and the Nominal Group Technique. *Organizational Behavior and Human Performance, 28,* 272-288.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes, 40,* 193-218.

Roth, P. L. (1993). Research trends in judgment and their implications for the Schmidt-Hunter global estimation procedure. *Organizational Behavior and Human Decision Processes, 54,* 299-319.

Russo, J. E., & Schoemaker, P. J. H. (1989). *Decision traps: Ten barriers to brilliant decision making and how to overcome them.* New York: Simon & Schuster.

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

Schwenk, C. R., & Cosier, R. A. (1980). Effects of the expert, devil's advocate, and dialectical inquiry methods on prediction performance. *Organizational Behavior and Human Decision Processes, 26,* 409-424.

Seaver, D. A. (1979). *Assessing probability with multiple individuals.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74,* 356-359.

Sniezek, J. A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting, 5,* 171-178.

Sniezek, J. A. (1992). Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes, 52,* 124-155.

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4,* 263-272.

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

Sniezek, J. A., Paese, P. W., & Furiya, S. (1990). *Dynamics of group discussion to consensus judgment: Disagreement and overconfidence.* Unpublished manuscript, University of Illinois at Urbana-Champaign.

Solomon, I. (1982). Probability assessment by individual auditors and audit teams: An empirical investigation. *Journal of Accounting Research, 20,* 689-710.

Stasser, G., Taylor, L. A., & Hanna, C. (1989). Information sampling in structured and unstructured discussions of three- and six-person groups. *Journal of Personality and Social Psychology, 57,* 67-78.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48,* 1467-1478.

Stasser, G., & Titus, W. (1987). Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology, 53,* 81–93.

Stephenson, G. M., Brandstätter, H., & Wagner, W. (1983). An experimental study of social performance and delay on the testimonial validity of story recall. *European Journal of Social Psychology, 13,* 175–191.

Stephenson, G. M., Clark, N. K., & Wade, G. S. (1986). Meetings make evidence? An experimental study of collaborative and individual recall of a simulated police interrogation. *Journal of Personality and Social Psychology, 50,* 1113–1122.

Stroebe, W., Diehl, M., & Abakoumkin, G. (1992). The illusion of group effectivity. *Personality and Social Psychology Bulletin, 18,* 643–650.

Tindale, R. S. (1983). *Individual and group reward allocation decisions in two situational contexts: The effects of relative need and performance.* Unpublished master's thesis, University of Illinois at Urbana-Champaign.

Tindale, R. S., Sheffey, S., & Filkins, J. (1990, November). *Conjunction errors by individuals and groups.* Paper presented at the annual meeting of the Society for Judgment and Decision Making, New Orleans, LA.

Trafimow, D., & Sniezek, J. A. (1994). Perceived expertise and its effect on confidence. *Organizational Behavior and Human Decision Processes, 57,* 290–302.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1130.

Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology, 58,* 582–592.

Van de Ven, A., & Delbecq, A. L. (1971). Nominal versus interacting group processes for committee decision-making effectiveness. *Academy of Management Journal, 14,* 203–212.

Wright, E. F., Lüüs, C. A. E., & Christie, S. D. (1990). Does group discussion facilitate the use of consensus information in making causal attributions? *Journal of Personality and Social Psychology, 59,* 261–269.

Wright, E. F., & Wells, G. L. (1985). Does group discussion attenuate the dispositional bias? *Journal of Applied Social Psychology, 15,* 531–546.

Wright, G., & Wisuda, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology, 23,* 219–224.

Yaniv, I., & Foster, D. P. (1992). *On graininess of judgment.* Unpublished manuscript.

Received May 16, 1994
Revision received August 5, 1994
Accepted September 29, 1994 ■