Psychological Review
2001, Vol. 108, No. 1, 183–203

Copyright 2001 by the American Psychological Association, Inc.
0033-295X/01/$5.00 DOI: 10.1037//0033-295X.108.1.183

Signal-Detection Analysis of Group Decision Making

Robert D. Sorkin, Christopher J. Hays, and Ryan West
University of Florida

How effectively can groups of people make yes-or-no decisions? To answer this question, we used signal-detection theory to model the behavior of groups of human participants in a visual detection task. The detection model specifies how performance depends on the group's size, the competence of the members, the correlation among members' judgments, the constraints on member interaction, and the group's decision rule. The model also allows specification of performance efficiency, which is a measure of how closely a group's performance matches the statistically optimal group. The performance of our groups was consistent with the theoretical predictions, but efficiency decreased as group size increased. This result was attributable to a decrease in the effort that members gave to their individual tasks rather than to an inefficiency in combining the information in the members' judgments.

How effectively can groups of people perform yes-or-no decision tasks, and how does their performance depend on the abilities of the individual members and the way they interact? We attempted to answer these questions by using signal-detection theory to model the behavior of groups of human participants in a visual detection task. The signal-detection model specifies how the accuracy of a group's performance depends on the group's size, the detection abilities of the individual members, the correlation among member judgments, the constraints on member interaction, and the group's decision rule. The model also allows specification of the efficiency of group performance; that is, it yields a measure of how closely the group's performance matches that of a hypothetical, statistically optimal group. This efficiency measure can be factored into separate components that describe how well the individual members performed their tasks and how effectively the group combined the information from the members into a group decision. The results of our experiments provide support for the signal-detection analysis and allow interesting conclusions to be made about the sources of inefficiency in the decision-making behavior of human groups.

Statistical arguments about the effects of group size and member competence on group performance have existed for more than 200 years, since Condorcet (1785) and, more recently, Einhorn, Hogarth, and Klempner (1977). According to the statistical argument, group performance should increase with group size, with the most rapid increase occurring when the competence of the group's members is high and when independent information is available to each member. These models assume that there is a statistically effective way to combine the members' judgments. If the expertise of the members varies within the group, each member's input should be weighted proportionally by the member's competence at the task (Grofman, Feld, & Owen, 1984; Grofman, Owen, & Feld, 1983; Nitzan & Paroush, 1982, 1984; Shapley & Grofman, 1984).

The empirical data on group performance indicate that human groups are generally less effective than would be predicted by statistical models that assume the optimal use of member information. In a fascinating sketch of 40 years of research on group decision making, Davis (1992) pointed out that most research has found group performance to be relatively inefficient. Group performance usually is superior to the average of individual performance but less than the statistical expectation (see also Hastie's, 1986, review). Moreover, many studies found that group performance either is insensitive to group size or that the advantage of size declines more rapidly than would be predicted from the statistical argument. All of these results can be attributed to inefficiencies in group function, such as might be caused by difficulties in member interaction or coordination, reduced member effort such as social loafing (Latané, Williams, & Harkins, 1979; Shepperd, 1993), or problems in combining judgments from multiple sources (Myung, Ramamoorti, & Bailey, 1996; Wallsten, Budescu, Erev, & Diederich, 1997).

Attempts to model group performance have used signal-detection theory (Batchelder & Romney, 1986; Erev, Gopher, Itkin, & Greenshpan, 1995; Metz & Shen, 1992; Pete, Pattipati, & Kleinman, 1993a, 1993b; Sorkin & Dai, 1994; Sorkin, West, & Robinson, 1998). In the group signal-detection situation, a group of observers is presented with an input that may have been either signal plus noise or noise alone. Each group member makes an observation and the group must then decide which of the two possible events gave rise to the input. Metz and Shen (1992) analyzed the gains in the detection accuracy of reading x-ray images that resulted from replicated readings by the same or multiple readers. Erev et al. (1995) examined the strategic interaction between two observers in a signal-detection task (i.e., when

Robert D. Sorkin, Christopher J. Hays, and Ryan West, Department of Psychology, University of Florida.

Christopher J. Hays is now with the U.S. Air Force.

This work was partially supported by grants from the Air Force Office of Scientific Research. Christopher J. Hays conducted Experiment 1 as part of the requirements for the master of science degree at the University of Florida. We thank Dr. John Tangney for his encouragement and support. Much of the article was written while Robert D. Sorkin was the Norman Munn Distinguished Scholar at The Flinders University of South Australia.

Correspondence concerning this article should be addressed to Robert D. Sorkin, Department of Psychology, University of Florida, P.O. Box 112250, Gainesville, Florida 32611. Electronic mail may be sent to sorkin@ufl.edu.

each observer's payoff structure was contingent on the outcome and the response of the other observer). Pete et al. (1993b) considered the case of multiple team members working in an uncertain, binary choice detection situation. They generalized the signal-detection model to consider the individuals' as well as the group's prior probability and payoff structure; that is, their model allowed joint optimization of the group aggregation rule and the individual decision rules of the group members.

Sorkin and Dai (1994) took a somewhat simpler approach to group signal detection than did Erev et al. (1995) and Pete et al. (1993b). Sorkin and Dai assumed that each group member could provide an estimate of the signal's likelihood of having occurred on a trial, and that the expertise of the members was known a priori to the group. These assumptions allowed them to sidestep the problem of how to aggregate binary responses from individuals who might have different biases toward the decision alternatives. Sorkin and Dai computed the performance accuracy that would result from the optimal aggregation of the members' likelihood estimates; this specified the performance of the ideal group. Later, we review the specific predictions of the ideal group analysis.

Because the performance level of the ideal group is the highest that may be achieved by any group, the ideal analysis specifies an upper bound on the performance that may be obtained from any group of human participants. Because the ideal model prescribes how the individual estimates of the group's members should be combined for a detection decision, the model also serves as a normative description of the behavior of human groups. For example, the ideal model assumes that each member of the group makes a continuous or graded estimate of the signal's likelihood, and that these estimates are then weighted by the member's detection ability. The weighted member estimates are then combined in an appropriate and noiseless fashion. We might expect that the decision-making process used by a group of human participants would violate some of these assumptions. Therefore, it is useful to consider the consequences of some specific (and perhaps drastic) violations of the ideal assumptions. For example, what are the performance consequences of requiring discrete rather than graded member judgments or of limiting the exchange of information among the members? The performance of specific suboptimal groups might define reasonable lower bounds on the performance to be expected from a group of human participants.

Consider a suboptimal group that arrives at a decision without any interaction or communication among the members. Suppose further that the group decision is determined by the aggregation of the members' unweighted binary (*yes-no*) votes; specifically, by application of a majority rule to the members' *yes* votes. We would expect that if a group used such a curtailed decision process, its performance would be well below the ideal level. Sorkin et al. (1998) used signal-detection theory to analyze the performance of such groups, known as Condorcet groups.[1] Condorcet groups are of interest because they provide an interesting kind of degenerate case of the optimal signal-detection group.

The inefficiency of a Condorcet group's performance is due to several factors. First, because there is no group interaction before voting, the group decision must be based on the unweighted combination of the members' decisions. Thus, information will be lost because the judgment of the least competent member counts as much as the judgment of the most competent member. Second, detailed information about the member estimates is lost because the member estimates are binary votes rather than graded judgments of signal likelihood. Additional potential losses occur because each member uses an independently determined criterion for making a binary yes response. Because members cannot use knowledge about other members' criteria, they cannot adjust (or readjust) their own response criterion for an optimal group setting.

Suboptimal models such as the Condorcet group may be useful for describing the behavior of some groups of human participants. This may be the case even when there are no externally imposed constraints on participant interaction or voting. That is, human groups may adopt aspects of the Condorcet decision mode even though more efficient modes of decision interaction are possible. Our initial hypothesis was that the upper bound on the performance of human groups would be given by the ideal group model, and that lower bounds on performance would be given by the Condorcet group model. (We assume that the members of the theoretical group have detection competencies equivalent to their counterparts in the human group.)

The psychophysical literature includes many studies that consider similar models, albeit in a different context (Green & Swets, 1966; Swets, 1984). The goal in many of these studies was to describe how a human observer aggregates stimulus information that arrives simultaneously or sequentially on multiple sources or on multiple channels. For an auditory task, these multiple sources might be different frequency bands or different earphone channels. For a visual task, these sources might be different spatial frequencies or different spatial positions. Can an observer perform this task with perfect efficiency? That is, can a person integrate all the relevant information that arrives on multiple channels (i.e., by performing the optimal statistical processing of the inputs as specified by an ideal signal-detection observer; see Green & Swets, 1966)? Alternatively, the observer's detection process may be suboptimal in a particular way. Perhaps the observations on different channels must be processed sequentially or first converted to separate binary (i.e., threshold) decisions, which are then combined.

Many of the multichannel psychophysical models are formally identical to putative models of group signal detection. Consider the following multichannel detection task. On a given trial, all of the multiple channels contain either noise alone or signal plus noise. The observer must observe all of the channels and make a single yes or no response to the possible occurrence of the (multiple-channel) signal. That is, the set of channel observations must be mapped to a *yes* or *no* response. The observations on the channels may or may not be correlated. The reader will see the similarity of this situation to the group detection case in which an array of multiple observers must monitor a single channel for the possible occurrence of either a noise-alone or signal-plus-noise condition on that channel. Each observer in the group makes an observation (possibly correlated) of the input, and the set of observations must be mapped to a single *yes-no* group decision.

Green and Swets (1966) and Swets (1984) discussed two generic classes of these psychophysical models: the observation-integration (OI) model and the decision-combination (DC) model. In the OI model, graded estimates of the signal's occurrence in

[1] For a broader definition of Condorcet-like groups, see Austen-Smith and Banks (1996).

each channel are available to be weighted and summed to form a final decision statistic. In the DC model, only binary responses are available from each channel, and these are combined for the final decision by applying a combination rule. To arrive at an overall *yes* decision, the combination rule could require (a) a single *yes* vote from any of the channels (the "union" rule), (b) *yes* votes from a specific majority of the channels, or (c) *yes* votes from all of the channels (the "intersection" rule). The channel signal-to-noise levels and the individual decision criteria used by each channel are important interacting variables in the DC model, and widely different performance can be obtained by changing the assumptions about their values.

It is clear that the different group models have counterparts in the multichannel psychophysical models and that the class of Condorcet models is equivalent to the DC models. When we discuss the results of our experiments, we make some further comparisons between these two classes of model. Durlach, Braido, and Ito (1986) reported a very elegant development of the OI class of psychophysical model. They developed a detailed model of the single-observer, multiple-channel auditory signal-detection situation, and their formulation provided the foundation for Sorkin and Dai's (1994) analysis of ideal group signal detection.

In this article we first describe the general detection task that is used in all of our experiments with human participants. Second, we provide a formal description of the group detection problem and of the ideal and Condorcet groups and briefly review their properties. Third, we report on experiments that assessed the detection performance of groups of human participants in different conditions. We compare the resulting performance to the predictions of the ideal and Condorcet models and argue that these models can account for much of the variance observed in the performance of the human participants. Finally, we report on a refined version of the experimental task that enables us to quantify the sources of inefficiency in group detection performance.

Signal-Detection Task

The basic task in our experiments was to judge whether the stimulus in an experimental trial was due to a signal-plus-noise or noise-alone condition. Participants were presented with a graphic display consisting of nine analog gauges similar to those shown in Figure 1, and they had to respond whether the display was due to a signal-plus-noise or noise-alone condition. The setting displayed on each gauge was generated from a normal distribution whose mean depended on the nature of the trial. On a signal-plus-noise trial, the settings on all of the gauges were drawn from the signal-plus-noise distribution, and on a noise-alone trial all the settings were drawn from the noise-alone distribution. The signal-plus-noise distribution had a higher mean than the noise-alone distribution, and both distributions had the same variance. In Figure 1, the means of the respective distributions are indicated by the labeled markers on the left-hand side of the display. Figure 1 illustrates a typical trial when a signal-plus-noise condition was present.

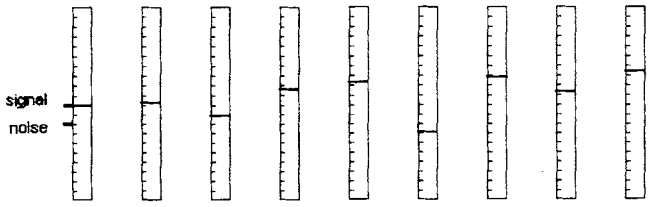

Figure 1. An example of the stimulus array presented to a participant on a signal-plus-noise trial of the experiment. On each trial, the values displayed on the nine gauges were drawn from either the signal-plus-noise or the noise-alone distribution. The thick ticks labeled "signal" and "noise" indicate, respectively, the means of these distributions. The value of the common standard deviation determined the difficulty of the task (see text).

Single-observer versions of this task have been studied extensively in our laboratory (Elvers & Sorkin, 1989; Montgomery & Sorkin, 1993, 1996; Sorkin, Mabry, Weldon, & Elvers, 1991; Sorkin, Robinson, & Berg, 1987). The difficulty of this task is determined by the display's physical and statistical parameters. The major physical parameters are the display duration and the visual angle subtended by the display. The statistical factors are the difference between the means of the signal-plus-noise and noise-alone distributions and the value of their standard deviation. If the physical parameters are fixed, the difficulty of a single-gauge display is directly proportional to the difference between the distribution means and inversely proportional to the standard deviation. If the nine gauge settings are generated independently, observer performance with the nine-element array should be $\sqrt{9}$ better than performance with a single gauge (Sorkin et al., 1991). This assumes that the information from all of the nine gauges is available to the observer, which is the case when the display duration is sufficiently long. Sorkin et al. (1991) studied the effects of the display duration, the size of the display, and the type of gauge used. They showed that short durations (less than 180 ms) prevent the observer from gaining information from gauges near the visual periphery. Their experiments also indicated that if the physical conditions are constant, most of the variance in an observer's performance is determined by the means and standard deviation of the gauge distributions.

In the present study, we tested both individual participants and groups of from 5 to 10 participants under different display and member interaction manipulations. Our experiments allowed group members to communicate their estimates of signal likelihood and did not impose constraints on the particular decision rule that the group used. After a group or individual decision was made on a trial, full feedback about the correct answer was provided to the participants. In certain group conditions, information about the responses of other participants was provided. In the single-participant conditions, the participant received a monetary payoff that depended on the accuracy of his or her performance. In the group task conditions, the monetary payoff to the participants depended on the accuracy of the group's detection performance. In the next section, we review the theoretical analyses of the ideal and Condorcet groups.

Group Signal-Detection Theory

An important benefit of applying signal-detection theory to a decision task is that it enables the experimenter to compute, from the obtained group or individual data, separate indices of performance accuracy (d') and bias (criterion or c). The accuracy measure, d', is expressed in standard deviate units. The d' index can vary between 0, for a chance level of performance, and approxi-

mately 4, for errorless performance. The criterion measure, c, is expressed in similar units. A value of c equal to 0 indicates that there is no preference toward a signal-plus-noise or noise-alone response, and a positive value indicates that there is a preference for the noise-alone response (Macmillan & Creelman, 1991). We used these measures to describe both individual and group performance in our experiments.

The general group signal-detection paradigm is shown in Figure 2. There are m members of the group. On each trial, the array of m members is presented either with a signal-plus-noise event or noise-alone event, and the group must decide which was presented. Each member has an individual index of detection accuracy, d'_i. On a signal-plus-noise trial, each member receives an input equal to μ_i, and on a noise-alone trial each member receives an input equal to 0. The task is made difficult by the presence of two Gaussian, zero-mean noise sources to each member, σ^2_{com} and σ^2_i. The first noise component, σ^2_{com}, is the variance of a noise source that is common to all the members, and the second, σ^2_i, is the variance of a noise source that is unique to each member. To arrive at a group decision, the members' judgments must be combined in some manner. In a specific decision situation, the members might express their judgments as binary responses (yes, no), continuous (graded) ratings of estimated signal likelihood or in other ways. The group decision process might include the exchange of information among the members about member likelihood estimates, confidence, and biases.

Ideal Group Model

An additional benefit of detection theory is that it enables one to specify the behavior of the statistically optimal or ideal detection system (Green & Swets, 1966; Tanner & Birdsall, 1958). By definition, an ideal detection system uses an optimal decision rule (one based on a likelihood ratio statistic) and suffers from no additional sources of noise or error. On average, an ideal detection system will produce the most accurate detection performance. The ideal analysis informs us about important task variables and how they may influence human performance.

Figure 3 shows how the general group signal-detection paradigm is modified to arrive at the ideal detection system of Sorkin and Dai (1994). They assumed that, although the unique noise source to each member is independent of the noise to any other member, the magnitude of the unique sources is constant across the array of members and is equal to σ^2_{ind}; that is, $\sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_{ind}$. Then each member's estimate, X_i, will be normally distributed with a mean of μ_i or 0 (respectively, depending on whether the trial was a signal-plus-noise or noise-alone trial) and with a variance equal to the sum of the common and unique noise variances. The index of detection sensitivity, d'_i, for an individual member is the difference between the means of the input on signal-plus-noise and noise-alone trials, divided by the square root of the total noise variance:

$$d'_i = \mu_i/(\sigma^2_{com} + \sigma^2_{ind})^{1/2}. \quad (1)$$

By definition, the correlation ρ between any pair of members is

$$\rho = \sigma^2_{com}/(\sigma^2_{com} + \sigma^2_{ind}). \quad (2)$$

Normalizing the total variance,

$$\sigma^2_{com} + \sigma^2_{ind} = 1. \quad (3)$$

Then,

$$d'_i = \mu_i,\ \sigma^2_{com} = \rho, \text{ and } \sigma^2_{int} = 1 - \rho. \quad (4)$$

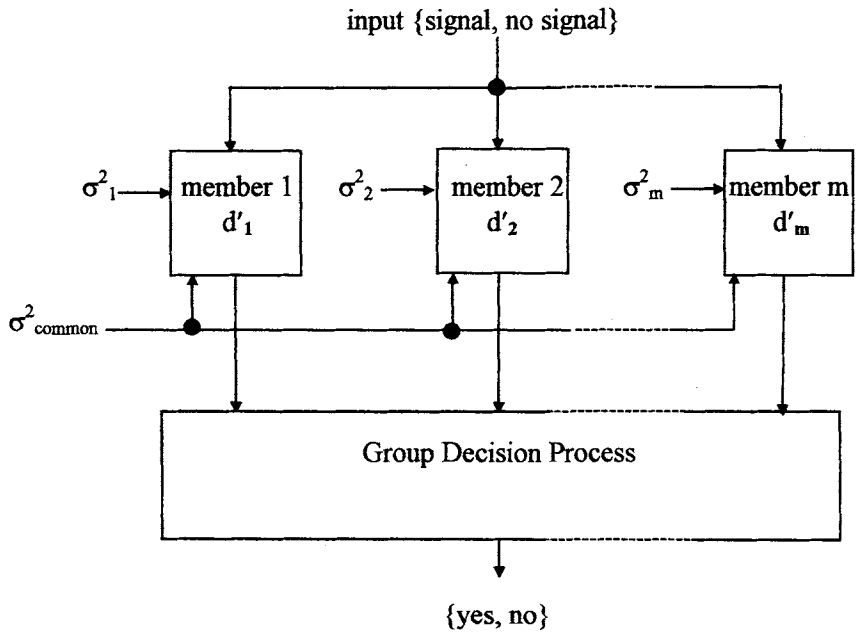

Figure 2. Diagram of a group signal-detection system composed of m members. Each member is subjected to two sources of Gaussian noise: one unique (σ^2_i) and one common (σ^2_{common}) to the other members. The member outputs are combined to form the group decision (see text).

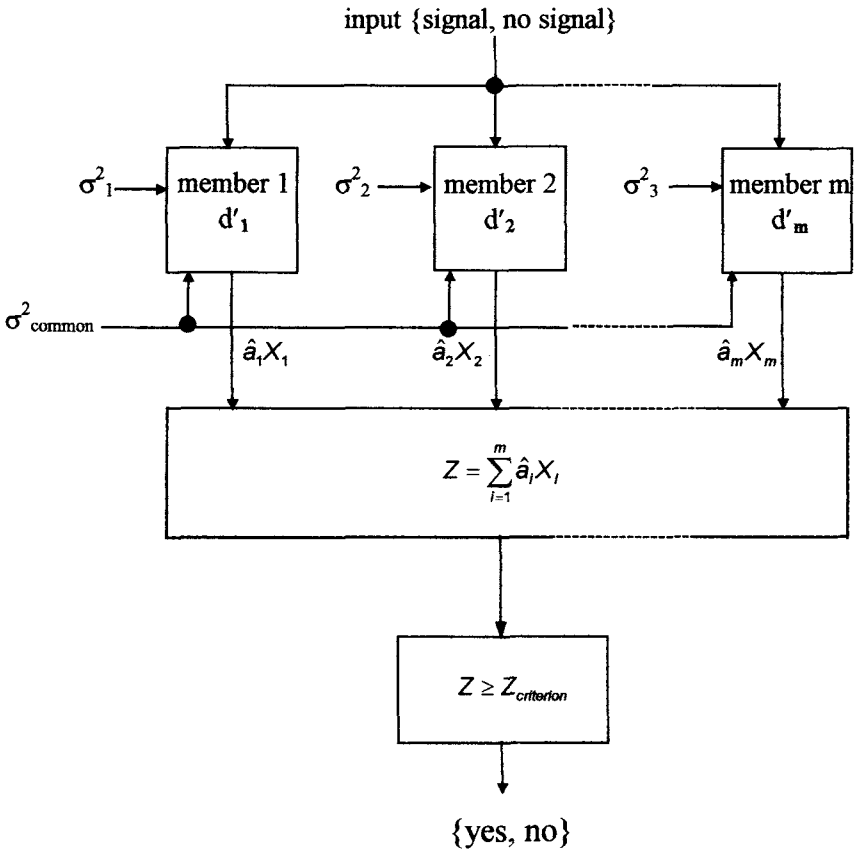

Figure 3. Diagram of an ideal group signal-detection system composed of m members (after Sorkin & Dai, 1994). Each member is subjected to two sources of Gaussian noise: one unique (σ_i^2) and one common (σ^2_{common}) to the other members. The decision variable, Z, is formed from the weighted sum of the member estimates (see text).

How should the member estimates be combined to make the signal-plus-noise/noise-alone decision? A decision statistic that is equivalent to a likelihood ratio statistic can be formed by summing the weighted estimates of the individual members (see, e.g., Ashby & Maddox, 1992; Berg & Green, 1990; Durlach et al., 1986; Green, 1992; Sorkin & Dai, 1994). The group decision statistic is

$$Z = \sum_{i=1}^{m} \hat{a}_i X_i, \tag{5}$$

where the $\{\hat{a}_i\}$ are optimal decision weights applied to the estimates of the individual members. To arrive at the group response on a trial, the aggregate judgment, Z, is compared with a criterion value, Z_c. When $Z >= Z_c$, the group response is *yes,* indicating a signal-plus-noise decision, and when $Z < Z_c$, the response is *no,* indicating noise-alone. The optimal weights $\{\hat{a}_i\}$ are specified (Durlach et al., 1986; Sorkin & Dai, 1994) by

$$\hat{a}_i = [1 + \rho(m - 1)]d'_i - m\,\rho\,\text{mean}(d'), \tag{6}$$

where mean(d') is the mean of the members' individual indices of detectability, d'_i. From Equation 6, it can be seen that the optimal weights are proportional to the individual indices of detectability. Therefore, the estimates of members having high d's should be afforded higher weights than members having small d's. Using the optimal weights yields the ideal group performance (Sorkin & Dai, 1994),

$$d'_{\text{ideal}} = \sqrt{m}\left[\frac{\text{Var}(d')}{1 - \rho} + \frac{[\text{mean}(d')]^2}{1 + \rho(m - 1)}\right]^{1/2}. \tag{7}$$

When the correlation is 0, Equation 7 reduces to the expression (Green & Swets, 1966):

$$d'_{\text{ideal}} = \left[\sum_{i=1}^{m} (d'_i)^2\right]^{1/2}. \tag{8}$$

Equation 7 specifies the maximum performance to be expected from a group of m members that have a specified mean, variance, and correlation. The equation also suggests what to expect from groups whose performance is similar to but less than the ideal's: (a) Group performance will increase when m increases; (b) performance will increase as $\sqrt{m}$ when $\rho = 0$; (c) performance will increase when the variance in member ability increases; and (d) much of the advantage of group size will be lost when $\rho > 0.25$.

The top curve of Figure 4 shows ideal group performance as a function of group size, for a group with $\rho = 0$ and the member parameters mean(d') = 0.78 and var(d') = 0.014. Although not shown on Figure 4, the ideal function flattens as ρ differs from 0 (see Sorkin & Dai, 1994). The ideal function defines the upper bound on the performance of any group that has the same member statistics.

What are the consequences of using nonoptimal weights when the members have different detection indices? If uniform weights are used (i.e., if $a_i = 1/m$), performance is given by the right-hand term of Equation 7:

$$d'_{\text{uniform}} = \frac{\text{mean}(d')\sqrt{m}}{\sqrt{1 + \rho(m-1)}}. \tag{9}$$

It follows that Equation 9 is the best performance predicted for a simple version of the Delphi group (Hillman, Hessel, Swensson, & Herman, 1977). For such groups, we assume that the members' judgments are not identified with individual members, and that the group decision is not informed about the detection competence of the individual members. The group thus lacks a basis for using anything other than uniform weights. Finally, the performance of a group that uses the set of arbitrary weights, $\{a_i\}$, is given by[2]

$$d'_{\text{weight}} = \frac{\sum_{i=1}^{m} a_i d'_i}{\sqrt{(1-\rho)\sum_{i=1}^{m} a_i^2 + \rho\left(\sum_{i=1}^{m} a_i\right)^2}}. \tag{10}$$

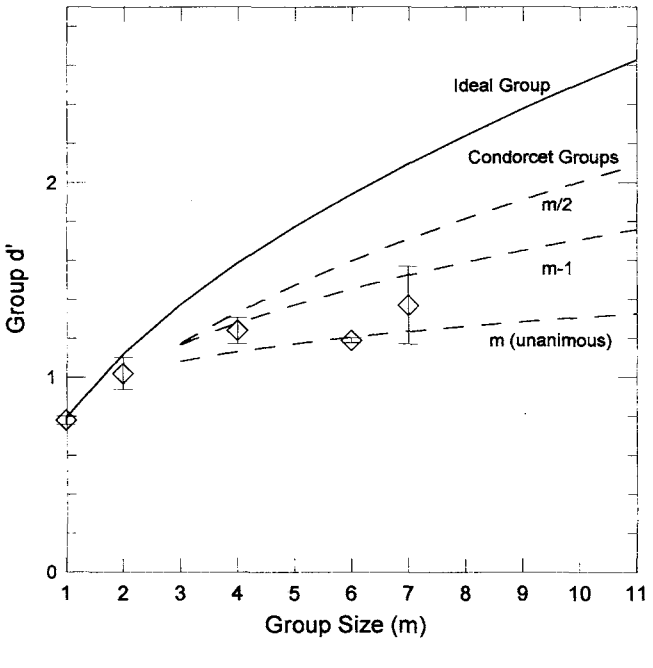

Figure 4. The performance (group detection indices) of the ideal and Condorcet groups is plotted as a function of group size. The solid line shows the performance of the ideal group with member parameters—mean(d') = 0.78, var(d') = 0.014—and assuming $\rho = 0$ and optimal weights. The dashed lines represent the performance of groups whose members do not interact and whose group decision is based on a majority of the (unweighted) binary votes of the members (see text). The group labeled *m*/2 requires *yes* votes from at least half of its members; the $m - 1$ group requires *yes* votes from all members but one; and the m group requires a unanimous *yes* vote. These groups have the same distribution of member sensitivities as the ideal group, and its members are neutrally biased ($c = 0$). The diamond symbols are the average group performance obtained from the uniform display signal-to-noise ratio = 1, $\rho = 0$ conditions (see text: Experiment 1, Results and Discussion). The brackets indicate ±1 *SEM*.

Condorcet Group Model

Figure 5 shows the member array for a Condorcet group. Sorkin et al. (1998) called this group a Condorcet group after the Marquis de Condorcet's (1785) analysis of similar groups. To arrive at detection theory predictions, Sorkin et al. assumed that the group's members do not interact with each other before voting, and that the members' only motivation is to maximize the payoff for correct group decisions. As in the ideal case, the group is composed of m members, and each group member is characterized by a detection sensitivity, d'_i. In addition, each member has a response criterion, c_i. Each member observes the stimulus input and makes an estimate of the likelihood that the input on that trial was caused by a signal-plus-noise event. This estimate is then compared with the member's response criterion, c_i, to make a binary judgment of yes or no. A single ballot is taken, and the group decision is determined by application of a majority rule to the binary votes of the members.

To calculate the performance of these groups, Sorkin et al. (1998) assumed values for the mean and standard deviation of the members' detection indices in a hypothetical detection situation. They generated receiver operating characteristic (ROC) curves for the group's behavior by varying the mean of the members' response criteria. The ROC is a plot of the hit probability (the probability of responding *yes* given signal plus noise) versus the false-alarm probability (the probability of responding *yes* given noise alone) at a given level of display difficulty. The ROC is the locus of all hit and false-alarm probabilities that are possible to achieve with a fixed detection accuracy; thus, the properties of the ROC determine the system's index of sensitivity.

The Sorkin et al. (1998) analysis showed that the Condorcet group ROCs resembled ideal ROCs in shape but were lower than the ideal curve and dependent on the particular majority rule that the group used. That is, Condorcet hit rates were lower than the ideal hit rates and the Condorcet false-alarm rates were higher than the ideal false-alarm rates. In addition, ROCs for Condorcet groups that used stringent majority rules, such as a three-quarters majority or a unanimous rule, were slightly distorted in shape and below the curves that used less stringent rules. As a consequence, the detection sensitivity of a Condorcet group is lower than the ideal function and is dependent on the particular majority rule used.

The dashed curves in Figure 4 show the predicted performance of Condorcet groups having the same individual detection properties as the ideal curve plotted. The lower two dashed curves specify the performance of the Condorcet group operating with more stringent majority decision rules. The highest dashed curve

[2] See Appendix for derivations of Equation 10.

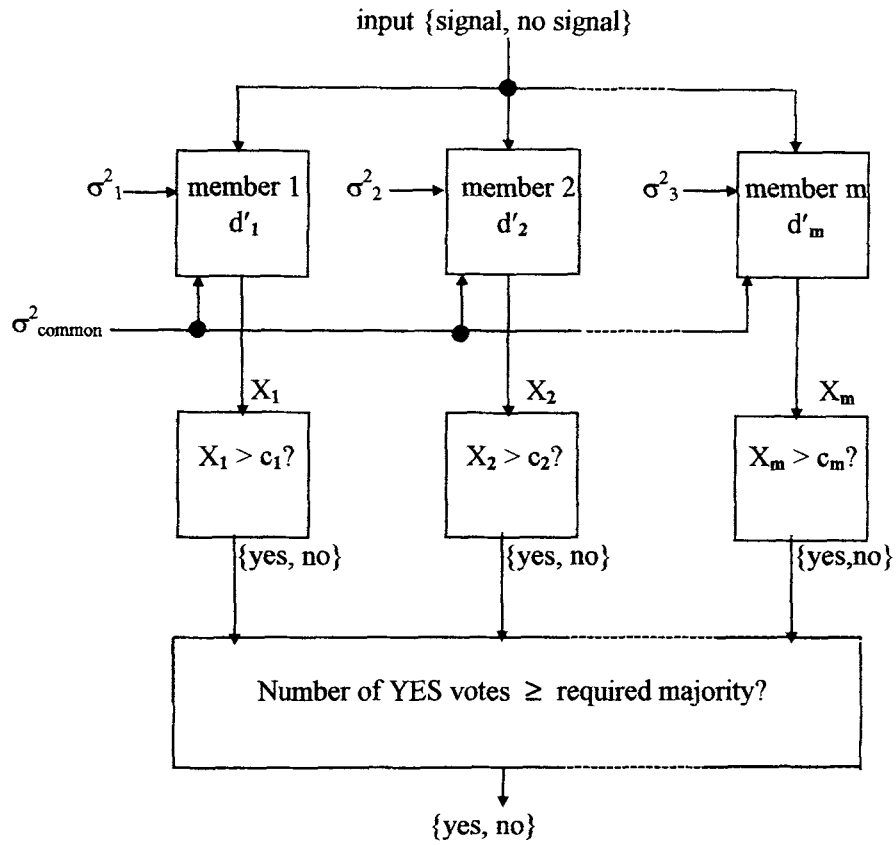

Figure 5. Diagram of a Condorcet group signal-detection system composed of m members (after Sorkin, West, & Robinson, 1998). The $\{d'_i\}$ are the member detection indices and the $\{c_i\}$ are the member response criteria. The decision is based on the majority rule of the members' binary votes (see text).

specifies the performance of the group that uses a simple majority rule (at least half of the members, $m/2$, must vote *yes*); the next lower dashed curve is produced by the more stringent rule (all but one member must vote *yes*); and the lowest dashed curve is produced by the unanimous rule (all members must vote *yes*). Nonzero levels of ρ produce qualitatively similar curves with smaller absolute differences between the performance of the different majority rules (not shown).

We have replotted the ideal and Condorcet curves on logarithmic coordinates in Figure 6, along with those for a higher average detection index of 1.5. Plotting these curves on log-log coordinates allows one to compare the growth rate of the functions with earlier psychophysical models. In the ideal case, the group detection sensitivity increases with group size with an exponent of 0.5, the square root. The OI model, of course, has a slope that is identical to the ideal. The detection index increases at a lesser rate for the Condorcet model; the slopes in the simple majority, $m - 1$, and unanimous cases are approximately 0.43, 0.31, and 0.16, respectively, and are approximately the same for the two different d' levels considered. Citing an analysis by T. Birdsall, Swets (1984) estimated that, at midlevel signals and midrange criterion values, the DC model produces slopes of about one third. These slopes are very close to the Condorcet, $m - 1$ case.

In addition to their theoretical analysis of the Condorcet group, Sorkin et al. (1998) asked several questions about the performance of groups of human observers under Condorcet-type task constraints. First, would one obtain the predicted decreases in the performance of the human groups as the majority rule was made more stringent? Second, would a more stringent rule produce a change in the behavior of the individual participants? A more stringent rule should cause the group hit and false-alarm rates to decrease, resulting in more conservative overall performance. The question was whether this would have an effect on the member's individual detection sensitivities or decision criteria. Specifically, would forcing the group decision to be more conservative cause the members to shift toward more liberal response criteria? Sorkin et al. (1998) ran groups of from 5 to 7 people in a visual signal-detection task in which the group members did not communicate with each other and the group decision was automatically determined by the majority rule of the binary votes of the members. The groups of human participants exhibited the same behavior as the model. Performance was best for the simple majority rule and worst for the unanimous rule. Two-third and three-quarter majority rules produced appropriately ordered intermediate levels of performance. Some participants adopted more liberal response criteria when the majority rule was more strict. Sorkin et al. (1998) noted that group members can reduce the effect of a strict majority rule on the group criterion by making their individual criteria more liberal, but they cannot undo the deleterious effect of a stricter majority rule on the group's performance accuracy.

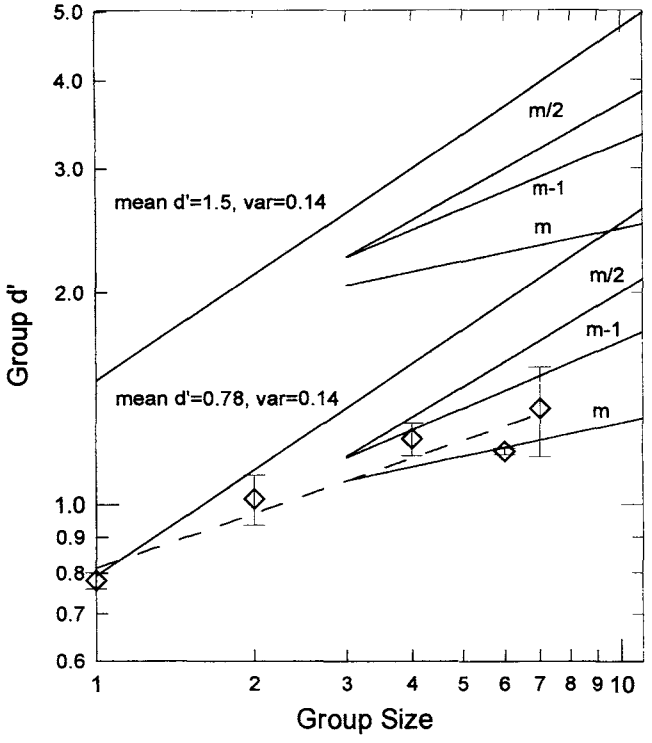

Figure 6. The performance (group detection indices) of the ideal and Condorcet groups is plotted as a function of group size for two levels of display difficulty (log-log coordinates). In each case the upper solid line shows the performance of the ideal group and the shorter lines show the Condorcet groups. The diamond symbols (and dashed line) show the average group performance from the uniform display signal-to-noise ratio = 1, $\rho = 0$ conditions (see text: Experiment 1, Results and Discussion). The brackets indicate ± 1 *SEM*.

Group Efficiency

It is useful to have a summary measure that describes how much the observed performance of a group of human observers differs from that of a hypothetical reference group such as the ideal group or the Condorcet group. The degree to which the performance of the real group is less than a reference optimal level is given by the efficiency measure, η (Tanner & Birdsall, 1958), where

$$\eta = \left(\frac{d'_{\text{observed}}}{d'_{\text{ideal}}} \right)^2 . \qquad (11)$$

Efficiency is defined as a ratio of squared d's, because in many sensory situations $(d'_{\text{ideal}})^2$ is proportional to signal energy. Thus, an efficiency of 0.60 means that an optimal detector could match the human's performance with a signal that contained only 60% of the energy needed by the human.

Obtaining a measure of a human group's efficiency depends on having an appropriate definition of the "ideal" reference group to be compared with the human group. The first piece of information needed to specify the reference group is the set of individual detection indices of the human group's members. In some experiments, it may be difficult to determine these indices from the members' or group's behavior. The second piece of information needed is the extent to which the observations of the human members are correlated, so that the appropriate correlation can be specified for the members of the reference group. Again, it may be difficult to determine the nature of this correlation without making additional assumptions or performing additional experiments. The last piece of information needed concerns the presence of any external constraints on the interaction and decision making of the human members in the task. For example, if group members were not allowed to interact and their decision was determined by a binary vote, it would be reasonable to define optimal performance by using the Condorcet rather than the ideal group model as a reference. Once the member indices, correlation among observations, and interaction constraints have been specified for the reference group, one can compare the performance of the human and reference groups and calculate a measure of the overall efficiency of the human group's performance.

Given that one has obtained a measure of overall efficiency, it may be possible to factor this measure into subordinate factors that describe different aspects of the group's performance. For example, one may wish to specify how much of the loss in overall detection efficiency is due to changes in the individual detection efforts of the members or to inefficiencies in combining the members' judgments into a group decision. Different members may make numerical judgments of signal likelihood in different ways, and this variability may lead to decreased accuracy in the group's decision. Perhaps the group gives inappropriate weight to the estimates from some members. Rather than weight each member's judgment in proportion to that member's expertise, the group may weight all members' judgments equally or even pathologically (e.g., by giving higher weights to the least competent or loudest members). Later, we develop a technique for quantifying these different sources of inefficiency.

In Experiment 1 of this study, we intentionally manipulated the correlation between member judgments by controlling the correlation between the stimuli presented to different members. This allowed us to examine the effects of member correlation on group performance. We also tested the possibility that the group's performance would be degraded by use of an inappropriate weighting strategy. To assess that possibility, we calculated the weights given to the judgments of individual members of different groups. In addition, we varied the difficulty of the detection task (i.e., the display signal-to-noise ratio) for each member of a group so that the detection performance of some of the group's members would be approximately twice that of the others. The group was able to sense this discrepancy in member competence and use appropriate weights in its decisions.

The performance of the larger groups in Experiment 1 suggested that they might have been operating under some restrictions in between-member interaction, possibly because of the way the experiment was conducted. Therefore, Experiment 2 was designed to optimize the effectiveness of between-member communication. We also measured detection efficiency much more accurately by assessing the detection effort of each member of the group while performing the group task. A major interest in Experiment 2 was to observe the effect of group size on efficiency. We were able to calculate a precise measure of overall efficiency and to separately quantify losses in efficiency caused by (a) changes in the individual detection efforts of the members and (b) the inefficient aggregation of member judgments.

Experiment 1

The Sorkin et al. (1998) study showed that the level of detection performance specified by the Condorcet model is very similar to that exhibited by groups of noninteracting human participants. The constraints on information sharing in this task impose a performance ceiling that is far below what could be achieved by an interacting ideal group. In Experiment 1 we removed any constraints on member interaction and tried to assess how much of an increase in group performance could be achieved by allowing the group to communicate their estimates of signal likelihood freely. We also investigated the effects of nonzero correlation among the judgments of the members, and we tested whether interacting groups could improve the efficiency of their decision making by using information about differences in the detection abilities of the individual group members.

Method

Observers performed a graphic signal-detection task, either individually or in groups of 5 or 7 members. On each trial, observers were presented with a multiple-element visual display consisting of nine analog gauges, similar to those shown in Figure 1. After the display, observers had to indicate whether the display represented a signal-plus-noise or noise-alone condition. The values displayed on the nine gauges were determined by sampling from one of two normal distributions: for signal, $\mu_S = 5$, and for noise-alone $\mu_n = 4$. The value of the common standard deviation, σ, was set equal to 1.5, 2, 2.5, or 3 in different conditions of the experiment.

Because the difference between the signal-plus-noise and noise-alone means was fixed, the magnitude of σ determined the display detectability. A high value for σ indicates that the detection task will be difficult. Consider the best detection performance that might be obtained with a single-element display, when $\sigma = 3$. Detection performance (d') based only on this element would be equal to $(\mu_s - \mu_n)/\sigma = 0.33$, which is a very low value. With nine independent display elements, the best detection performance would be $\sqrt{9}$ times the element detectability (i.e., 0.33 $\sqrt{9} = 1.0$). In this article, we use the value of $3(\mu_s - \mu_n)/\sigma = 3/\sigma$ to specify the display signal-to-noise ratio (DSNR) to characterize the detectability of different experimental conditions. Thus, element standard deviation values of 3, 2.5, 2, and 1.5 yield DSNR conditions of 1, 1.2, 1.5, and 2, respectively.

Normally, we would not expect that a human observer would achieve a d' of 2 when detecting a signal presented at a DSNR equal to 2. In experiments using similar graphical materials, Sorkin et al. (1991) and Montgomery and Sorkin (1996) showed that limitations on human processing, primarily caused by the display's brief duration, resulted in performance that was about 75% to 80% of the predicted values (i.e., η between 0.56 and 0.64).

Participants. Eight University of Florida students, seven women and one man, participated in the study. All of the participants had normal or corrected-to-normal visual acuity. Participants were paid $4.25 per hour plus an incentive bonus that was based on the accuracy of performance. In the individual conditions, the bonus depended on the accuracy of the individual's performance. In the group conditions, the bonus depended on the accuracy of the group's performance. In Experiment 1 the bonus averaged approximately $0.40 per person per hour.

Apparatus and stimuli. Stimulus generation and presentation were done with Insight 4086-33 computers arranged in a small local area network, synchronized by a separate 4086 computer. The stimuli were displayed on 14-in. CTX color monitors (1024 × 768 SVGA monitor, 72-Hz refresh rate at 640 × 480 resolution). The monitors were set for maximum contrast, with the intensity at approximately 100 cd/m^2 measured from a uniform field. Participants sat approximately 27 in. away from the monitor in a quiet, fluorescent-lit laboratory room; the nine gauges subtended a visual angle of approximately 8 degrees vertical by 22 degrees horizontal. Responses were made on a standard computer keyboard. During the group phase of the experiment, participants were seated close to each other, but they could not see monitors other than their own.

The individual display elements, shown in Figure 1, consisted of two parallel vertical lines with tick marks on the left line, dividing the gauge into 20 intervals. A value of 0 was represented at the bottom of the gauge and a value of 10 at the top. Two larger tick marks on the leftmost gauge marked the mean of the noise and signal-plus-noise distributions. On a given trial, all of the elements displayed values that had been drawn, independently, from the same distribution. Half of the trials (randomly) were drawn from the signal and half from the noise distribution. Stimulus duration was 370 ms during practice sessions and 320 ms during all experimental trials. A centered cross (0.5 in.) fixation stimulus (200 ms) preceded the stimulus, and a white masking screen (200 ms) followed presentation of the stimulus. After the masker, the screen was blank for a short time period before the response. After the response, feedback about the correctness of the response was given.

Procedure. Each participant was first tested alone in the individual detection sessions. The individual sessions were run before any group conditions and then rerun again after all the group conditions. A trial block consisted of 125 trials at a given DSNR level. An experimental session consisted of 16 blocks, presented in sequences of four blocks at a given DSNR. The DSNR levels were randomized both within and across sessions. After two practice sessions, participants performed the task for 2,000 trials at each of the four DSNR conditions. A session took approximately 1.5 hr, and participants were encouraged to take rest breaks after each block.

After the individual session, participants were tested in groups of 2, 4, 6, and 7 members. Individuals were randomly assigned to one group of 6 or 7 members, two groups of 4 members, or two groups of 2 members. The 2-member groups were randomly chosen from the 4-member groups. However, the male subject was purposely excluded from the 2-member groups to minimize any chance that his presence would bias that group's decision (Clement & Schiereck, 1973). We had planned to use 8 individuals in the largest group. As a result of absenteeism and scheduling problems, the large groups tested consisted of only 6 or 7, rather than 8, members.

The trial blocks in the group sessions consisted of 100 trials, run in two four-block sets, randomized within and across sessions with respect to DSNR and correlation. The trial procedure was the same as for the individual sessions, except for important differences in the group response procedure. In the individual sessions, the participant had up to 1,000 ms after the masking screen to make a response. In the group sessions, a 700-ms blank screen was presented after the masking screen. After this screen, one member of the group was randomly selected and received a screen message telling her that she was to give the group's answer. There was no time limit for a response. The other members of the group did not make any manual response. Group members were encouraged to talk about their judgments both during the 700-ms blank period and the period after the group responder had been selected. Usually, members began announcing their estimates as soon as the display ended. These were usually in the form of a short binary statement, such as "I think signal" or "I'm sure signal," from each of the members. After the responder entered the response, each member received feedback about the nature of the trial and the correctness of the response.

In addition to changing DSNR, two additional manipulations were made during the group sessions. First, the distribution of DSNR was set either the same for all members of the group (the equal DSNR condition) or, for some 4-person groups, intentionally varied so that the task difficulty for two members was twice that for the others (the unequal DSNR condition). Second, the stimulus displays were either independent ($\rho = 0$ condition) for all group members or were partially correlated ($\rho = 0.25$) between

group members. Table 1 summarizes the conditions for the group sessions. (Some of the low-difficulty conditions were omitted for the larger groups, because the group performance would have been too high to measure accurately.)

The correlation between group members was manipulated using a method described by Jeffress and Robinson (1962) and Sorkin (1990) in auditory experiments. The method can be understood by considering how the values were generated for Element 1 (the left-most element) in Participant A's and Participant B's displays. In the independent condition (ρ = 0), each of the elements was drawn from a separate normal distribution as follows: For Participant A, the value of Element 1 was equal to x_a; for Participant B, Element 1 was equal to x_b, where x_a and x_b were normally distributed, independent, zero-mean, equal variance, random variables. However, in the correlated condition (ρ = 0.25), the value for each participant's display element was generated as follows: For Participant A, the value of Element 1 was set equal to $0.87x_a + 0.5x_c$. For Participant B, Element 1 was set equal to $0.87x_b + 0.5x_c$, where x_a, x_b, and x_c were independent, normal, zero-mean, equal variance, random variables (i.e., the principle is the same as stated by Equation 2). In the correlated condition, the corresponding elements in all pairs of displays were generated in a similar fashion.

Results and Discussion

We evaluated performance in all conditions of the experiment by calculating detection indices (d') and criterion (c) measures based on the obtained individual and group hit and false-alarm rates (Macmillan & Creelman, 1991). The criterion measures were generally near zero, indicating an absence of response bias, and they did not vary consistently across conditions, participants, or groups.[3] Therefore, our analysis consider (a) the individual and group detection indices and (b) the weighting strategies used in the group conditions.

Table 1
Experimental Conditions for the Group Sessions of Experiment 1

Size (m)	Correlation (ρ)	DSNR	Group membership	Number of blocks of 100 trials
2	0	1	S1 S2	12
2	0	1	S5 S6	8
2	0	2	S1 S2	8
2	0	2	S5 S6	8
2	0.25	1	S1 S2	8
2	0.25	1	S5 S6	8
2	0.25	2	S1 S2	8
2	0.25	2	S5 S6	8
4	0	1	S1 S2 S3 S4	8
4	0	1	S5 S6 S7 S8	8
4	0	1.5	S1 S2 S3 S4	8
4	0	1.5	S5 S6 S7 S8	12
4	0.25	1	S1 S2 S3 S4	8
4	0.25	1	S5 S6 S7 S8	8
4	0.25	1.5	S1 S2 S3 S4	8
4	0.25	1.5	S5 S6 S7 S8	11
4U	0	1, 1, 2, 2	S2 S5 S7 S8	8
4U	0	1, 2, 1, 2	S2 S5 S7 S8	4
4U	0	2, 2, 1, 1	S2 S5 S7 S8	4
6	0	1	S1 S3 S5 S6 S7 S8	4
7	0	1	S2 S3 S4 S5 S6 S7 S8	4
6	0	1	S2 S3 S4 S6 S7 S8	4
6	0	1.2	S1 S3 S5 S6 S7 S8	4
7	0	1.2	S2 S3 S4 S5 S6 S7 S8	4
7	0	1.2	S1 S2 S3 S4 S5 S6 S7	4
7	0.25	1	S1 S2 S3 S5 S6 S7 S8	4
7	0.25	1	S1 S2 S3 S4 S5 S6 S7	4
7	0.25	1.2	S2 S3 S4 S5 S6 S7 S8	4
6	0.25	1.2	S2 S3 S4 S6 S7 S8	4

Note. DSNR = display signal-to-noise ratio.

Figure 7 is a plot of group performance as a function of the DSNR for all groups in which the DSNR was equal for each member. The dashed curve on Panels A and B (1 symbols) in Figure 7 shows the average detection performance of individuals as a function of DSNR. Individual performance was essentially a linear function of DSNR, consistent with the predictions of traditional, single-observer, detection theory. The absolute level of performance also was consistent with previous results using a similar task (Montgomery & Sorkin, 1996; Sorkin et al., 1991). Average individual detectability at a DNSR equal to 1 was 0.78 with a standard deviation of 0.12. There was no significant difference between individual performance in the test (pregroup sessions) and retest (postgroup sessions) conditions.

The indices of observer performance obtained in the individual conditions can be converted to measures of individual detection efficiency (Equation 11). Observer efficiency in the individual sessions averaged 0.61, was moderately consistent across participants (the largest standard deviation in η_i across participants at any DSNR was 0.17), and was highly consistent across DSNRs (the largest standard deviation in η_i across conditions for any participant was 0.1). These efficiencies are consistent with those obtained in our previous single-observer experiments.

The solid lines in Figure 7 show the effect of DSNR on the performance of groups of 2, 4, 6, and 7 members (indicated by the plotted symbols 2, 4, 6, 7); the left and right panels of the figure, respectively, show the results for the uncorrelated and correlated conditions. As in the individual case, performance increased with DSNR (the increase with DSNR for 6- and 7-member groups did not reach statistical significance). Group performance increased with group size; for the ρ = 0 condition, $F(4, 51) = 4.56$, $p < .003$. The small decrease in performance in the ρ = 0.25 condition was consistent with Equation 7.

The nature of the increase in group performance with group size is evident in Figures 4 and 6; the average group detection indices (diamond symbols) are plotted as a function of group size for all the zero correlation conditions having a DSNR of 1.0. The figures also show the predictions of the ideal and Condorcet models that assume groups having the same detection properties as the participant groups. The performance of the human groups increased with group size, but at a rate that was less than either the ideal model or the simple majority Condorcet model. In Figure 6, the best fitting line (on log-log) coordinates has a slope of 0.26 and is positioned

[3] The average criterion in the group conditions was 0.021 with a standard deviation of 0.17. The lack of criterion effects was consistent with our previously reported data on individual participants in similar tasks (Sorkin et al., 1991) and generally consistent with results reported by Pete et al. (1993a) in their study of distributed detection by 3-person groups. The criteria used by Pete et al.'s observers (operating points) reflected a relatively neutral bias and were somewhat insensitive to experimental manipulation of event probability and cost structure. However, the direction of the observers' criterion shifts was in the direction of the optimal criterion. In all of our experimental conditions, the optimal group criterion was 0 (c = 0).

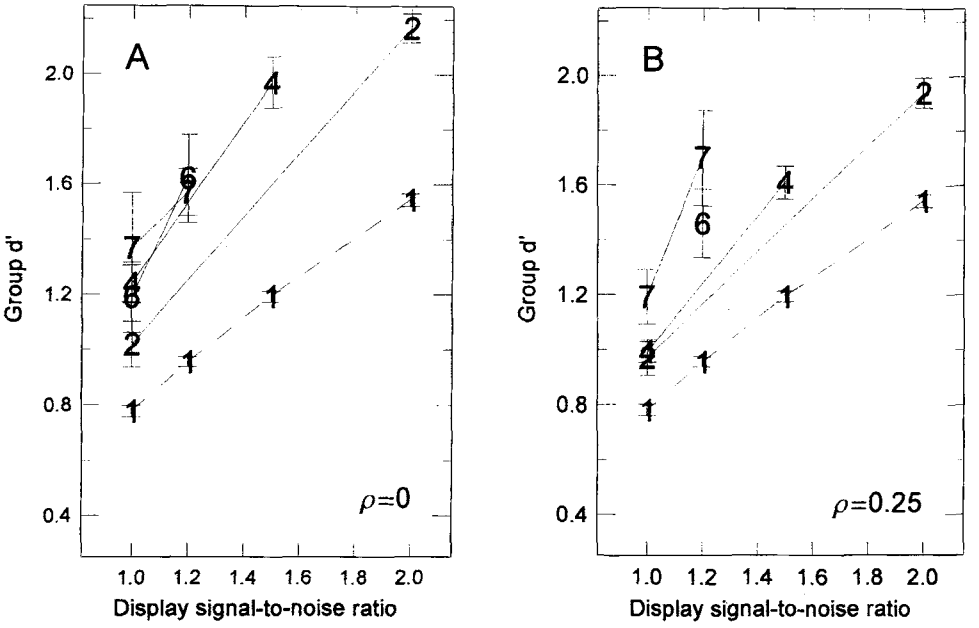

Figure 7. The obtained performance (d') of all groups, plotted as a function of the display signal-to-noise ratio. Panels A and B show, respectively, the data for the $\rho = 0$ and $\rho = 0.25$ correlation condition. The plotted symbols (1, 2, 4, 6, 7) indicate the group size. The dashed line shows the data for the individual (1) condition (repeated in each panel). The brackets indicate $\pm SEM$.

above the Condorcet-unanimous function and just below the Condorcet-($m - 1$) function. It is clear that the overall efficiency of human group performance decreased with group size.

We calculated the overall efficiency of detection performance in the group conditions by using an ideal reference group whose individual members had the same detection indices as obtained from our human participants in their individual detection sessions. That is, the d' indices obtained in the members' individual sessions were taken to specify the putative values for those members in the group conditions. For example, to calculate the ideal performance of the 4-person group that consisted of Participants S1, S2, S5, and S6, the d' values that had been obtained in the individual sessions for these individuals were used to evaluate Equation 7. The correlation among group members for the ideal group was assumed to be the experimental value.

This calculation of group efficiency was made for all the groups in Experiment 1. As in the individual participant conditions, efficiency in the group conditions was highly consistent across different DSNR levels and somewhat consistent across groups of similar size. The solid line and circle symbols in Figure 8 show the obtained group efficiency measures as a function of group size, for the $\rho = 0$ conditions. The figure shows clearly that overall efficiency started out at a very high level, 90% for the 2-person groups, but fell rapidly as group size was increased. In the next section we consider the possibility that the decrease in efficiency was due to the use of a nonoptimal weighting strategy.

Efficiency of decision weights. A correlational technique developed by Lutfi (1995) and Richards and Zhu (1994) was used to determine the relative weights assigned to each member during the response deliberation process. This technique is based on Berg's (1989, 1990) conditional on single-stimulus (COSS) analysis of decision weights in individual sensory tasks. Sorkin et al. (1991) utilized the COSS technique in a study of visual displays using stimuli similar to those in the present experiment. The basic idea

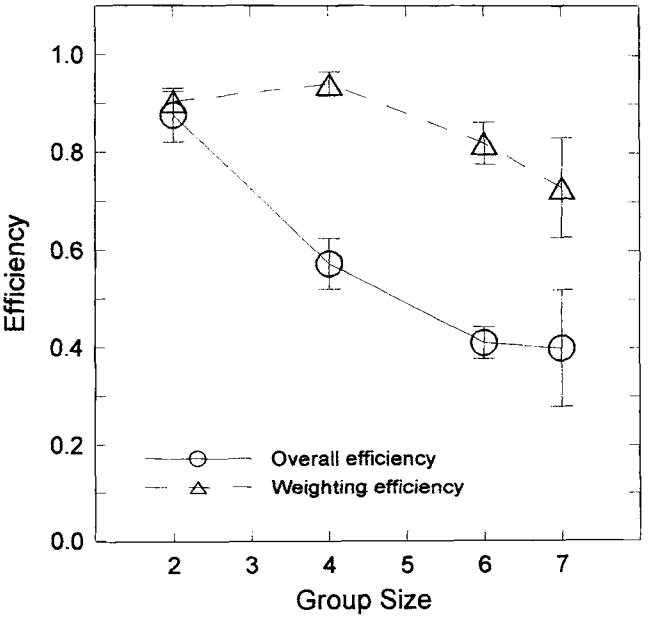

Figure 8. A plot of performance efficiency as a function of group size for the $\rho = 0$ conditions (circles). The triangles (and dashed line) show the weighting efficiency, the group's ability to weight a member's contribution by the member's expertise.

of the correlational analysis is to compute the point-biserial correlation between the stimulus presented to each observer and the group's response over trials. Because we had recorded the mean value of the nine-element display presented to each observer on each trial, we could compute the correlation between this value and the group's response. This correlation (scaled by the variance of the stimulus data and using partial correlations when $\rho > 0$) provides a measure of the relative impact of that observer's stimulus on the group's decision. In previous applications of this method, the goal was to assess the relative influence of different components of the stimulus on the response of a single observer. In the present experiment, we assessed the relative influence of different observers on the response of the group.

Using the correlational technique, we estimated the weights given to each observer in each condition. In some of the equal DSNR conditions (6 person, $\rho = 0$, DSNR = 1.2; 7 person, $\rho = 0.25$, DSNR = 1; 4 person, $\rho = 0$, DSNR = 1; and 4 person, $\rho = 0.25$, DSNR = 1.5), the magnitude of the obtained decision weights had approximately the same ordering as the ideal weights. That is, the largest weights were given to the most sensitive observers. In other conditions (7 person, $\rho = 0.25$, DSNR = 1.2; 7 person, $\rho = 0$, DSNR = 1.2; 4 person, $\rho = 0$, DSNR = 1.5; and 4 person, $\rho = 0.25$, DSNR = 1), the ordering appeared to be random. However, in most cases, the variation in the obtained decision weights for a condition was approximately the same as the variation in the ideal weights for that condition. In only one case (7 person, $\rho = 0$, DSNR = 1.2) was a negative weight given to an observer. The particular observer was late for two consecutive experimental sessions, thereby forcing the group to wait before being able to start the session. Apparently, this behavior resulted in her being given a negative weight.[4] Figure 9 shows the consistent weighting pattern that was obtained in the unequal DSNR conditions (4 person, $\rho = 0$). Here, the weights corresponded closely to the ideal values. In all three conditions, the two highest weights were given to the two members with the highest d's (and highest DSNRs).

We estimated that the standard deviation, σ_a, of the obtained weights was approximately 0.04 (based on a statistical argument and a separate computer simulation that produced essentially the same result). Using that estimate, 24 of the 29 conditions tested produced weights that were within a 99% confidence interval around the set of optimal weights (i.e., 5 of 29 weights differed significantly from the ideal). This result is not very surprising, given the small variation in the ideal weights for the equal DSNR conditions. This variation was due entirely to the variance in the participants' d's; at a constant DSNR = 1, $\sigma_{d'} = 0.12$. In the unequal DSNR condition, where $\sigma_{d'} > 0.45$, none of the unequal DSNR conditions produced weights that were significantly different from the ideal.

Berg (1990) coined the term weighting efficiency (η_{weight}) to describe how accurately weights were assigned in a detection task. Weighting efficiency is equal to

$$\eta_{\text{weight}} = \left(\frac{d'_{\text{actual-weights}}}{d'_{\text{optimal-weights}}} \right)^2, \tag{12}$$

where $d'_{\text{actual-weights}}$ is the index of detectability that would have been obtained if the group's only inefficiency was due to using the obtained, rather than the ideal, weights, and $d'_{\text{optimal-weights}}$ is the index that would have been obtained using optimal weights. This index is calculated by entering Equation 10 with the d'_i indices obtained in the individual sessions and the weights derived from the correlational analysis of the group session (and the calculated optimal weights). The dashed line and triangle symbols on Figure 8 show the obtained weighting efficiencies and allow comparison with the overall detection efficiencies (circle symbols). Although there was an apparent drop in weighting efficiency with group size, the decrease cannot account for the large decrease in overall efficiency with size. In Experiment 2 we extended Berg's analysis to the general group decision situation, and we partitioned the overall efficiency measure into several additional sources of efficiency in the group's decision process.

The analysis of weighting efficiency indicated that the groups were effective at weighting the judgments of individual members according to the members' detection competence. Appropriate weighting by member d' was very clear in the unequal DSNR conditions and also was evident in many of the equal difficulty conditions, even though there was no strong payoff consequence for using an optimal weighting strategy in the uniform DSNR conditions. We should point out that it is not necessary for the group to make an accurate estimate of the d'_i of every other member because most of the time the group members will convey confidence information along with their judgments. A member having a high-difficulty display will find it difficult to make signal-plus-noise/noise-alone discriminations and will, if honest, communicate that uncertainty to the group. For example, a member who has a low-difficulty display may be very emphatic about conveying her judgment. Furthermore, a group soon would lose confidence in, and hence lower the weights given to, a member whose estimates were consistently at odds with the trial-by-trial feedback.

The preceding weights analysis may not have revealed one nonoptimal weighting effect that could have decreased overall efficiency. Suppose that the person designated to make the group response consistently gave a higher weight to her own judgment, and that every member followed that same strategy.[5] Because the choice of responder was random and the higher weight for each responder would be averaged across every member who responded, this higher-weight-to-responder strategy would be hidden from our weighting analysis. We tested for this strategy by comparing the correlation between a member's stimulus and the group's response when that member was the responder and when that member was not the responder.

The data from the 4-member equal DSNR ($\rho = 0$) condition were partitioned into the trials when a member was the designated responder and the trials when that member was not the designated responder. We then calculated the correlation between the group

[4] It is interesting that the group did not simply ignore this individual's input by giving it a zero weight. Instead, they "punished" her by making group responses that were contrary to her advice. Alternatively, she may have responded to social ostracism by intentionally providing bad estimates.

[5] Harvey and Harries (1999) found that people put more weight on forecasts that were their own (whether or not they were labeled as such) or were labeled as their own (when they were not) than on forecasts that were neither their own nor labeled as their own. They discussed several factors that may be involved in the overweighting of one's own estimates.

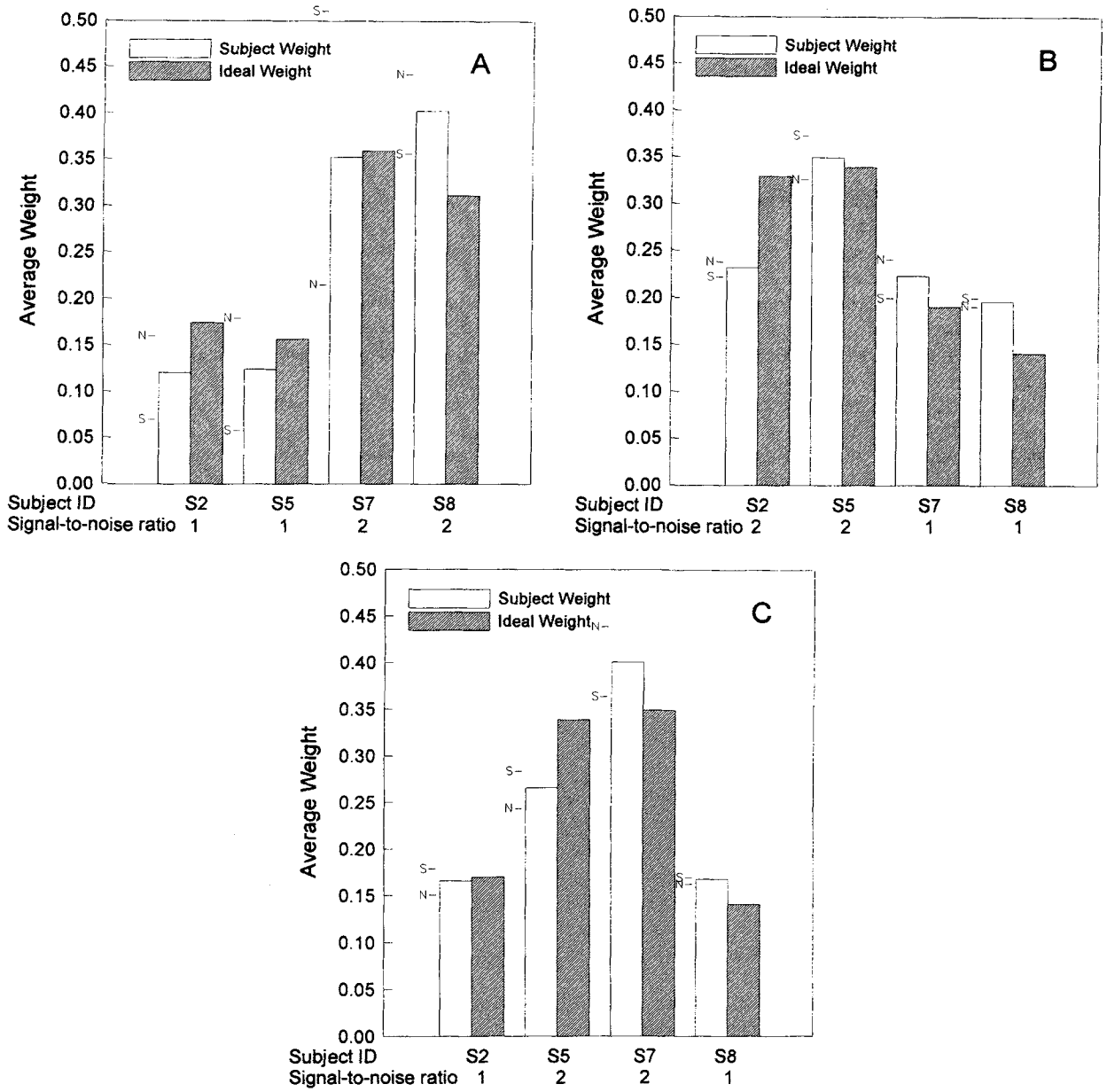

Figure 9. The average relative weights for each member of the 4-member groups in the unequal difficulty conditions. The *S* and *N* symbols show the weights calculated separately on signal-plus-noise and noise-alone trials. The ordering of obtained weights is close to the ordering of ideal weights.

response and that member's stimulus (separately) for the two sets of partitioned trials. Two drawbacks of this analysis are the very small number of trials on which the weight is computed and the necessity to compare weights computed from different sets of trials. However, the results were unambiguous: weights were consistently higher when a participant was responder than when not. The size of the responder effect, in terms of average difference in decision weight, was approximately 0.15.

How much of a drop in efficiency would be produced by a responder effect of this magnitude? Using Equation 10, we estimated the drop in performance that would result from a consistent weight increment of 0.15 given to the responder over the other members. The uniform weight case was used as a comparison, because the effect was assumed to be averaged across all group members. In all cases, the estimated decrease in efficiency (for equal DSNR and $\rho = 0$) was less than 0.10. This decrement is much less than the observed drop in efficiency with group size. Therefore, we reject inefficiency in assigning decision weights as a major source of group detection inefficiency in this task.

Possible causes of decreased efficiency with size. One possible cause of the decreased detection efficiency that we observed is that the participants used a noninteractive, binary voting strategy such

as used by a Condorcet group. It is difficult to test this hypothesis directly. When plotted as a function of group size, the data do resemble the $(m - 1)$ Condorcet function. We monitored the interactions of our groups and noted that many groups took binary ballots during their deliberations. However, it was apparent that members often communicated graded likelihood information when they conveyed their binary votes; that is, they varied the tone of their voices from tentative to emphatic and they included descriptive phrases such as "I think," "definitely a signal," and "not sure." Furthermore, the observed performance was well below that of the Condorcet simple majority rule, and there is no reason to expect that the group would have used a more stringent majority rule.

We believe that there are two likely causes and one unlikely cause for the observed decrease in group efficiency with size. The first likely cause is that as the group size was increased, the group members were more rushed, less complete in their deliberations, and effectively more Condorcet like. This possibility is supported by the fact that there was a small incentive for completing the deliberations rapidly. We did not record the average deliberation time in the experiment, but we know that it took much longer to complete a 100-trial block with the larger groups. If participants were attempting to finish the same number of trial blocks per session—to maximize their per hour pay—they would have worked more hurriedly in the larger groups. To remedy this shortcoming in Experiment 2, we attempted to make the number of trials per hour independent of group size.

The second possible explanation for the efficiency decrease is simply that individual members may have worked less hard and decreased their detection effort as a function of the group size. Because of the statistical advantage of aggregating observations, even a very small decrement in the detection index for the individual members will produce a moderate to large decrement in group performance. Such small decrements would be very difficult to observe without running a large number of trials. Furthermore, we did not have a precise estimate of the detection index for individual members during the group sessions. Recall that to calculate group efficiency we used the individual detection indices that had been obtained in separate, individual detection sessions with each participant. Thus, we had no way of knowing whether the individual detection indices varied in different group size conditions.

Finally, an unlikely explanation for the observed decrease in group efficiency is that the correlation between member observations was actually not zero as set in the experiment. Perhaps the judgments of members was correlated at some small but significant level. Experiment 2 was designed to answer all these questions.

Experiment 2

The decision task in Experiment 1 had at least two weaknesses. First, it placed a slight premium on making decisions rapidly, and this may have led the group deliberations to decrease in effectiveness as the groups increased in size. Second, it did not require the individual members to make a formal signal-noise response on each trial. We did not require individual responses because we wanted to minimize the chance that members would have a strong commitment to a particular *yes* or *no* decision before the group's deliberation on each trial. We thought that the absence of mandatory individual decisions would increase the likelihood that members would contribute graded estimates of signal likelihood to the group discussion. We also thought that members would be more open to the influence of other members' opinions if they had not committed themselves to a binary decision. This turned out to be a weakness in the experiment's design, because, without any formal response from a member, we could not accurately calculate the level of detection performance for members in different conditions. If a member's detection effort changed as a consequence of some aspect of the group test situation, such as the group's size, we could not easily detect that result. As a consequence, we were unable to attribute the observed loss in group efficiency with size to a particular cause. However, we were reasonably confident in ruling out losses in efficiency resulting from inappropriate weighting strategies.

These design limitations were addressed in Experiment 2. First, we revised the task contingencies so that there was no financial advantage to completing a block of trials in a shorter period of time. Second, we required each group member to make a numerical estimate of the likelihood of signal occurrence on each trial. These estimates were recorded and displayed to every member of the group so that they could be used during the group deliberations. These estimates could be checked to determine the actual correlation between the judgments of any pair of group members. As in Experiment 1, we also recorded the mean value of the nine-element stimulus display that was presented to each member on each trial. These two pieces of information were used to calculate measures of virtual performance in each condition: (a) the best (ideal) group performance that was possible in the condition, and (b) the best performance that would be possible from a hypothetical group that used the information available in its members' estimates of signal likelihood. In the next sections, we show how these virtual measures of performance enabled us to calculate the component efficiencies of the group's decision-making performance.

Method

Participants. Twenty-one people from the university and community (11 females and 10 males) participated as paid volunteers. Participants were recruited by advertising in the school newspaper and from undergraduate psychology classes. As in Experiment 1, all participants were paid $4.25 an hour plus a bonus for correct responses, but the bonus in Experiment 2 was approximately $1 per hour. Before testing, participants were given extensive training to ensure task competency and minimize learning effects during the experiment.

Procedure. The basic task, apparatus, and stimuli were the same as in Experiment 1, except that the stimulus duration was 160 ms. After presentation of the stimulus display and before group deliberation, each member was required to provide a rating, using a horizontal slider, of the estimated likelihood of signal occurrence on that trial. After all of a group's members had generated a rating, the ratings were rank ordered and displayed graphically to all members of the group. The ratings were presented in a way designed to facilitate their use during the group decision process. Figure 10 shows a sample display of the ratings displayed to members of the group before deliberation. The ratings were ordered from the smallest (noise alone most likely) to the largest (signal plus noise most likely) and presented in a graphic display so that each member could quickly estimate the degree of consensus favoring the noise-alone and signal-plus-noise hypotheses. After the onset of this display, one of the members was randomly selected to respond with the group decision. There was no time

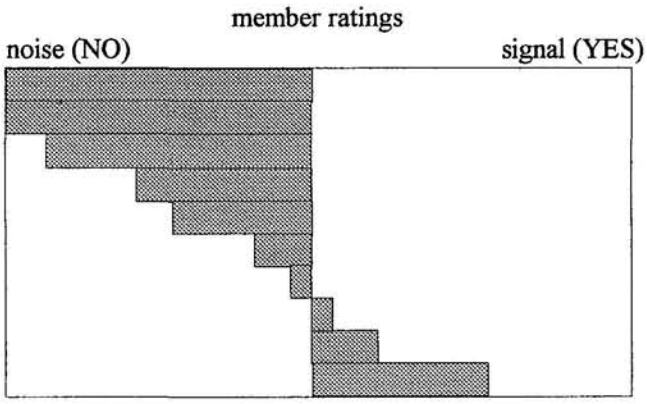

Figure 10. An example of the display of member estimates of signal likelihood on a trial. The members' estimates were displayed from top to bottom in order of their favoring the likelihood of signal-plus-noise on that trial.

limit to the deliberation process. The number of trial blocks run in each group size condition and the number of trial blocks per experimental session were fixed. Participants knew that there was adequate time in the session for completing the required number of trial blocks. As in Experiment 1, the incentive payoff to the individual members was determined only by the accuracy of the group's decision. Conversation among group members was allowed and occasionally occurred after the likelihood estimates were displayed and until the presentation of the fixation stimulus on the next trial.

A benefit of the rating display was that the group had an immediate indication whenever a respondent made a response that appeared to be inconsistent with the members' ratings. Such responses, although very rare, often were noted by verbal comment or complaint from group members. Because the graphic display ordered the responses in terms of rating magnitude, it may have been difficult for the group members to identify ratings with particular members. In most cases, however, the authorship of an apparently deviant rating was often either volunteered or otherwise evident by the end of the trial.

Table 2 summarizes the groups and conditions run. There were two 8-person groups. All groups were run for a total of 200 trials and at a display correlation equal to 0 except for the 4-person group, which was run for 300 trials. The DSNR was set at 1.33 except in the 10-member condition and in one of the 8-member conditions, where it was set at 1. This was done to avoid obtaining excessively high group detection indices ($d'_{group} > 3$) from the large groups.

Results and Discussion

Table 3 summarizes the obtained group performance averaged over each size group (3, 4, 5, 8, and 10 members). The first column in the table shows the group size and the second shows the DSNR used. The third column shows the values of DSNR $\sqrt{m}$ for each condition. These values predict the detection index for a hypothetical ideal group of m members, based on Equation 8, the group size, and the DSNR parameters. As in Experiment 1, the obtained group performance d'_{group} was based on the *yes-no* responses of the randomly designated responder for the group. The resulting performance is shown in column 4. In general, the obtained group performance increased with group size and DSNR. An analysis of the ratings of signal likelihood provided no evidence of a correlation between any two members' ratings or of a nonuniform

Table 2
Experimental Conditions and Group Membership for Experiment 2

Size (m)	Correlation (ρ)	DSNR	Group membership	Number of 100-trial blocks
3	0	1.33	S1 S2 S3	2
4	0	1.33	S4 S5 S6 S7	3
5	0	1.33	S1 S2 S3 S8 S9	2
8	0	1.33	S2 S3 S4 S5 S6 S7 S10 S11	
8	0	1	S12 S13 S14 S15 S16 S17 S18 S19	2
10	0	1	S12 S13 S14 S15 S16 S17 S18 S19 S20 S21	2

Note. DSNR = display signal-to-noise ratio.

correlation between members' ratings and the group response. Thus, there was no evidence of a correlation between member judgments or for the use of nonuniform weights.

Although the DSNR$\sqrt{m}$ values provide an estimate of optimal group performance, an improved estimate can be obtained from the actual sequence of stimuli presented to the groups of participants during the experiment. We determined this optimal level of performance by assuming a virtual ideal group; this group bases its decisions on a statistic, X_j, that is the sum of the actual stimuli presented to the participants on each experimental trial:

$$X_j = \sum_m x_{ij}, \tag{13}$$

where x_{ij} is the mean of the nine-element display presented to member i on trial j. X_j is formed by summing the x_{ij} values over the m members. Because the x_{ij} values contain all of the relevant information in the members' displays, the X_j statistic summarizes all of the information available on a trial. Thus, the theoretical best performance obtainable from a group that used all of the information in its displays is given by the performance of this statistic.

To specify the performance of the virtual statistic X_j, we calculated the detection index d'_X based on X_j. On each trial we summed the recorded mean displays and compared the value of that statistic

Table 3
Indices of Group Detection Performance in Experiment 2 as a Function of Group Size and Display Signal-to-Noise Ratio (DSNR)

Size (m)	DSNR	DSNR$\sqrt{m}$	d'_{group}	d'_x	d'_R	$d'_{AX'}$
3	1.33	2.30	1.47	2.00	1.69	1.59
4	1.33	2.66	1.99	2.94	2.19	2.61
5	1.33	2.97	1.67	2.32	1.59	1.99
8	1.33	3.76	2.55	4.12	2.50	2.50
8	1	2.83	1.06	2.33	1.28	1.58
10	1	3.16	1.42	2.79	1.55	1.62

Note. The d'_{group} index is based on the accuracy of the human group's decision on each trial. The d'_X index would be expected from a statistically optimal group that observed the same stimuli as the human group. The d'_R and $d'_{AX'}$ indices are component detection indices based on the members' ratings (see text).

to a criterion value, X_{crit}. If $X_j > X_{crit}$, the virtual group was considered to have made a signal-plus-noise response; otherwise, the group was considered to have made a noise-alone response. A virtual hit was scored if the statistic exceeded the criterion on a signal-plus-noise trial, and a virtual false alarm was scored if the statistic exceeded the criterion on a noise-alone trial. Thus, this procedure yielded a hit and false-alarm rate for a block of trials of the experiment. We repeated the procedure using a range of values for X_{crit}. The resulting indices were averaged to arrive at the virtual measure of performance d'_X shown in column 5 of the table.[6] All our subsequent calculations of efficiency will use d'_X as the estimate of d'_{ideal}. Note that if we had used displays with different signal-to-noise levels for each member, it would have been appropriate to use a weighted rather than an unweighted version of the X_j statistic:

$$AX_j = \sum_m \hat{a}_i x_{ij}, \tag{14}$$

where $\{\hat{a}_i\}$ is the set of optimal weights based on each member's DSNR.

To calculate the overall efficiency of the obtained group performance levels, we calculated the squared ratio of the obtained group and ideal detection indices:

$$\eta_{overall} = \left(\frac{d'_{group}}{d'_{ideal}}\right)^2 = \left(\frac{d'_{group}}{d'_X}\right)^2. \tag{15}$$

These computed efficiency values are shown in column 2 of Table 4 and are also plotted as a function of group size (the solid curve and circle symbols) in Figure 11. (The two 8-member conditions have been combined in Figure 11.) As in Experiment 1, there was a significant drop in overall group efficiency with size, $F(1, 11) = 9.27$, $p < .012$. Because of the limited number of 100-trial blocks (2, 3, or 4) on which to base an estimate of the standard error of the efficiency measure, we ran Monte Carlo simulations based on the obtained values of d'_X and d'_{group}. From each d' value, we generated a set of hits and false alarms and computed the squared ratio of the resulting d' indices. The simulation allowed an improved estimate of the standard deviation of the efficiency measures. These values are shown as the error brackets in Figure 11 and are approximately equal to or larger than those based on the actual trial blocks.

A major objective of Experiment 2 was to factor overall group efficiency into separate efficiency components that characterize the individual detection efforts of the members and the efficiency of the groups' aggregation of member information. In the next several paragraphs, we show how this partitioning was accomplished and how one can perform a more detailed analysis of the component efficiencies.

Recall that on each trial we had recorded the group's response and the mean of each member's individual display. In addition to those measures, we recorded the estimate of signal likelihood, r_{ij}, made by member i on trial j. We wished to calculate how much of the information in the aggregated member judgments was preserved in the group's decision. Suppose that the only information available to the group on a trial is the set of member ratings $\{r_{ij}\}$. A virtual statistic based on these ratings would provide a measure of the performance based on this information. We formed the statistic

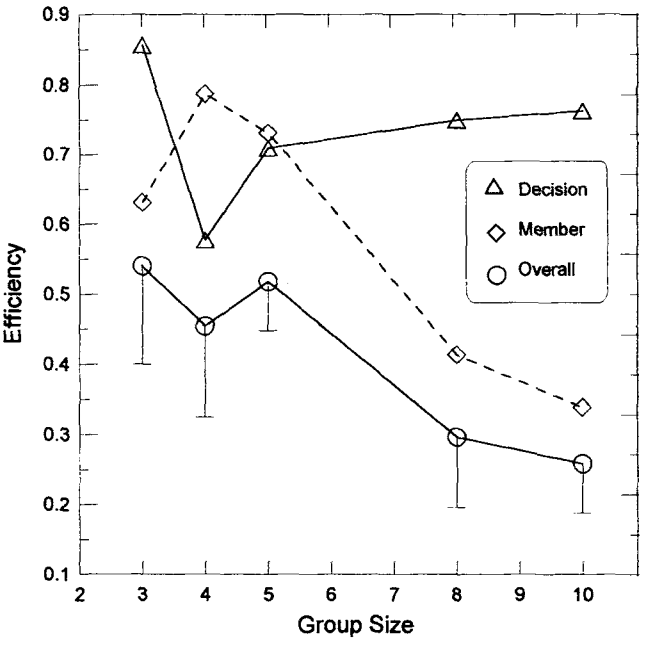

Figure 11. A plot of the efficiency of group performance as a function of group size. The lower solid line (circles) shows the average overall efficiency of the groups in Experiment 2. The brackets indicate ±1 *SEM*. The upper solid line (triangles) indicates the component efficiency associated with the group decision process. The dashed line (diamonds) indicates the component efficiency associated with the individual detection effort of the group's members. (See text.)

$$R_j = \sum_m r_{ij}, \tag{16}$$

where R_j is the summed rating of signal likelihood obtained from the members on trial j. This statistic was then evaluated in a manner similar to that used with the X_j statistic. That is, we calculated a virtual detection index based on R_j in the following way: A response from R_j was defined as a signal-plus-noise response if and only if $R_j > R_{crit}$. The resulting virtual hit and false-alarm rates were used to calculate the ratings-based index, d'_R, shown in column 6 of Table 3.

The ratio $(d'_{group}/d'_R)^2$ provides a crude measure of how efficiently the members' raw estimates $\{r_{ij}\}$ were combined to form the group's decision. If the obtained d'_{group} and d'_R indices were equal, we would conclude that all of the information in the members' ratings had been incorporated into the group decision. In fact,

[6] There are a variety of techniques to calculate a d' index from such rating data (e.g., find the best fitting ROC to all the points, estimate the area under the ROC; see Swets, 1996). In our experiment, both the display values and the rating values were essentially normally distributed, and only the middle criterion values produced nonzero false-alarm rates and nonunity hit rates. In view of the size of the trial block and the well-behaved nature of the (display and rating) values, a more computationally extensive procedure was deemed unnecessary. Note also that the displays from the m members were weighted uniformly in our calculations. If the members had been given displays with different signal-to-noise ratios, it would have been necessary to weight each x_{ij} by the member's DSNR.

the R_j statistic may not contain all of the information present in the members' ratings. For example, if there are differences in the individual indices of member detection ability (or in the members' DSNR), an improved statistic, AR_j, would use the appropriately weighted ratings on each trial,

$$AR_j = \sum_m \hat{a}_i r_{ij}, \tag{17}$$

where $\{\hat{a}_i\}$ is the set of optimal weights for the group based on the detection indices of the individual members.

A statistic that improves on AR_j follows from the assumption that there is variability in the way individual members make their rating judgments. Perhaps members use different scales for mapping their judgments of signal likelihood onto a graphical, numerical, or verbal output. If that were the case, performance could be greatly improved by normalizing the rating from each member before combining the ratings into a group decision. Recall that we recorded the mean display input x_{ij} and the resulting rating r_{ij} that was generated by each member during a block of trials. We can perform a linear regression on these data to predict the value of the display input for a member given that member's rating on a trial. The regression prediction enables us to generate estimates that have scaling properties that are consistent across the members. We define the predicted mean display input to participant i on trial j as x'_{ij} to distinguish it from the actual mean display input, x_{ij}. The value x'_{ij} is determined by r_{ij} and the regression coefficients for member j. We then define an optimally weighted statistic, AX'_j,

$$AX'_j = \sum_m \hat{a}_i x'_{ij}, \tag{18}$$

where $\{\hat{a}_i\}$ is the set of optimal weights for the group and x'_{ij} is the predicted display input for member i based on a linear regression of member i's rating r_{ij} on x_{ij}.

Now we can calculate the detection index, $d'_{AX'}$, for the virtual group that uses the AX'_j statistic. This detection index was computed in the same manner as for d'_X and d'_R. That is, a response from the AX'_j statistic was defined as a signal-plus-noise response if and only if $AX'_j > AX'_{\text{crit}}$. The resulting virtual hit and false-alarm rates were used to calculate the ratings-based index, $d'_{AX'}$. This detection index is especially useful because the AX'_j statistic contains all of the information in the aggregated judgments of the group's members. Regardless of how the group may use the information that it obtains from its members, on average it cannot do better than this index (column 7 of Table 3). To calculate the efficiency with which the group has used this information, we define the efficiency of the group's decision aggregation process as

$$\eta_{\text{decision}} = \left(\frac{d'_{\text{group}}}{d'_{AX'}}\right)^2. \tag{19}$$

This efficiency is the (squared) ratio of the actual group performance divided by the performance possible if the group had used all of the information available in the members' judgments.

Note that because the $d'_{AX'}$ index specifies all of the information that the members have obtained from the displayed stimuli, we can characterize the efficiency of the members' aggregate individual detection effort as

$$\eta_{\text{member}} = \left(\frac{d'_{AX'}}{d'_{\text{ideal}}}\right)^2. \tag{20}$$

This efficiency is the (squared) ratio of the performance possible based on all the information in the members' judgments divided by the performance possible given all of the information in the members' stimuli. Thus, η_{member} specifies the overall efficiency with which the members converted the stimulus information into rating judgments.

We can now decompose the overall group efficiency into its components. First, notice that we have partitioned the group efficiency into two major components:

$$\eta_{\text{group}} = \eta_{\text{decision}} \cdot \eta_{\text{member}} \tag{21}$$

because

$$\eta_{\text{group}} = \left(\frac{d'_{\text{group}}}{d'_{\text{ideal}}}\right)^2 = \left(\frac{d'_{\text{group}}}{d'_{AX'}}\right)^2 \cdot \left(\frac{d'_{AX'}}{d'_{\text{ideal}}}\right)^2. \tag{22}$$

These two efficiency components are shown in columns 3 and 4 of Table 4 and are also plotted as the upper solid line (triangle symbols) and dashed line (diamond symbols), respectively, in Figure 11. One can observe from Figure 11 that there is a clear distinction between the efficiency of the group decision process and the efficiency of the aggregate detection efforts of the individual members. Member effort declines with group size, whereas group decision efficiency does not.

We can define further efficiencies by factoring the group decision component into three subcomponents:

$$\eta_{\text{decision}} = \left(\frac{d'_{\text{group}}}{d'_{AX'}}\right)^2 = \left(\frac{d'_{\text{group}}}{d'_R}\right)^2 \cdot \left(\frac{d'_R}{d'_{AR}}\right)^2 \cdot \left(\frac{d'_{AR}}{d'_{AX'}}\right)^2. \tag{23}$$

The three terms on the right-hand side of Equation 23 may be described, from left to right respectively, as (a) the efficiency of decision based on ratings alone, (b) the weighting efficiency (as defined earlier in Equation 12), and (c) the efficiency arising out of consistency in the members' likelihood rating responses. That is, the latter term would be unity if there were no variability in the way that members generated their ratings.

Given the results of Experiments 1 and 2 (and the use of equal DSNR values in Experiment 2), we have assumed that the weighting efficiency was unity. Equation 23 may then be simplified to the ratings-alone and ratings-consistency components:

Table 4
Measures of Efficiency as a Function of Group Size

Size (m)	η_{group}	η_{decision}	η_{member}	$\eta_{\text{ratings-alone}}$	$\eta_{\text{ratings-consistency}}$
3	0.54	0.86	0.63	0.76	1.13
4	0.45	0.58	0.79	0.82	0.70
5	0.52	0.71	0.73	1.10	0.64
8	0.38	1.04	0.37	1.04	1.00
8	0.21	0.45	0.46	0.69	0.66
10	0.26	0.76	0.34	0.84	0.91

Note. Column 2 shows the overall efficiency of the groups' performance. Columns 3 and 4 partition the overall efficiency into two factors: η_{decision}, which characterizes the efficiency of aggregating member judgments, and η_{member}, which characterizes the efficiency of member detection. Columns 5 and 6 are subfactors of decision efficiency: the ratings-alone-efficiency and the ratings-consistency-efficiency (see text).

$$\eta_{\text{decision}} = \left(\frac{d'_{\text{group}}}{d'_{AX'}}\right)^2 = \left(\frac{d'_{\text{group}}}{d'_R}\right)^2 \cdot \left(\frac{d'_R}{d'_{AX'}}\right)^2. \quad (24)$$

Columns 5 and 6, respectively, in Table 4 summarize these decision subcomponents. Both subcomponents are relatively high and appear to be independent of group size.

General Discussion

A quantitative model of group decision making should be able to specify how performance depends on the formal properties of the task and on the abilities of the group members. A signal-detection analysis of the group decision task promises to meet these requirements. This analysis shows how the group decision rule, the constraints on member interaction, and the level and distribution of member expertise affect the accuracy of group performance. Furthermore, empirical tests of the theory lead naturally to hypotheses about the sources of inefficiency in the behavior of human groups and to estimates of the magnitude of those inefficiencies.

According to the detection theory analysis, the accuracy of a group's performance will be limited by the number and ability of the members and the correlation among member judgments. It is reasonable to expect that the performance of a human group will fall between two extremes: the performance of the statistically ideal detection group at the top end and the Condorcet (zero-interaction, binary-voting) detection group at the bottom end. We assume that there is a shared incentive for group members to work toward accurate group performance, and that no individual incentives conflict with the group incentive. We also assume that there are no constraints on member interaction that would cause the group to limit its mode of communication or deliberation. Finally, we assume that no members of the group exert an improper influence over the behavior of other members. Specifically, we assume that some members may have a greater influence on the group's decision than other members only if they are demonstrably more competent at the task.

We applied the detection analysis to a visual task in which groups had to make binary, diagnostic decisions based on the relatively brief display of noisy, graphically presented information. On the surface, this appears to be a very simple task. However, even when the task is performed by an individual participant, it involves many judgmental factors, including estimation, criterion setting, and the recall of information from memory. Perhaps in part because of its apparent simplicity, this task is a highly useful way to study information processing in groups of participants and falls well within the definition of group information processing as defined by Hinsz, Tindale, and Vollrath (1997).

The present experiments were designed to give groups of individuals the opportunity to reach their maximum level of performance in this decision task, by training, feedback, monetary payoff, and the opportunity for full member interaction. The results of Experiment 1 were generally consistent with the detection theory analysis in that (a) the level of group performance increased with group size, (b) the advantage of size decreased when member judgments were correlated, and (c) the group weighed the judgments of individual members approximately in accordance with the members' expertise. Libby, Trotman, and Zimmer (1987) showed that the variance in expertise within a group and the group's ability to recognize the relative expertise of its members are crucial in determining group performance when member interaction is allowed. Our results are consistent with their observations and with studies that show that groups can recognize the relative expertise of group members. For example, Henry (1993) demonstrated that groups can estimate the ability of members, even when no specific feedback about the correctness of judgments is provided.

Experiment 1 indicated that the efficiency of group performance decreased as the group size was increased or, equivalently, that the advantage of size declined more rapidly than the statistical expectation. This is consistent with studies that have found decision performance to be less than the statistical optimum (see Davis, 1992) and could be attributed to several possible causes, such as (a) nonzero correlations between member judgments, (b) inefficiencies in how the member judgments were combined to form a decision, and (c) decreases in member detection efforts with increased group size. In Experiment 2, we attempted to either control or assess the contribution of these different factors by requiring each group member to make an overt estimate of the likelihood that a signal-plus-noise event had occurred on each trial and by displaying these estimates to the group during its deliberation. Experiment 2 also enabled us to measure how efficiently the group used the judgments received from its members and to quantify the aggregate information losses resulting from decreased member effort.

Implicit in our analysis is the assumption that each member makes an observation of the stimulus, and that this observation leads to an internal, graded estimate of signal likelihood. One member's estimate could be correlated with another member's estimate as a consequence of common genetics, background, or experience or because the experimenter has intentionally manipulated the members' stimuli. This correlation between members is confined by assumption to the individual's perceptual stage of processing; that is, it is present earlier than the group decision stage of the system shown in Figures 2 and 3. Conversely, the influence of one member on another is confined to the decision stage of the system, specifically, the setting of decision weights on each member's contribution to the group decision (e.g., Kriss, Kinchla, & Darley, 1977; Robinson & Sorkin, 1985).

In Experiment 1 we attempted to control the intermember correlation by manipulating the stimulus information that was presented to each member. Our manipulation was designed to produce member judgments that were either independent or at a set correlation of 0.25. We observed a large drop in performance as a consequence of experimentally increasing the correlation from $\rho = 0$ to $\rho = 0.25$. If the member judgments had been correlated to a significant degree before our manipulation, the predicted effect of the increase in experimental correlation would have been much smaller (i.e., the drop it produced by shifting from 0.2 to 0.45 would have been very small). Therefore, we concluded that the effective correlation between members in the zero-correlation condition must have been near zero. This conclusion was confirmed by Experiment 2, in which we recorded the member judgments and could directly test the magnitude of the correlation between the member's estimates. We found that the average correlation between member estimates (considered within noise-alone and signal-plus-noise trials) was essentially 0.

For the reasons of member background and experience cited previously, one should not normally expect that the intermember correlation will be 0, particularly when the stimuli are not random

graphical patterns, as in the present study. Hinsz (1990) reported on a group recognition memory experiment using a signal-detection methodology. Hinsz's experiment involved 6 participants having the member properties: mean(d') = 1.04 and var(d') = 0.19. He obtained a group of d' of 2.11 in his experiment. This is much higher than the average member d' of 1.04 but less than the ideal value of 2.77. What accounts for the decreased efficiency of his participants? The discrepancy between ideal and actual group performance is probably not attributable to inappropriate weighting of the members' judgments, because a uniform weighting strategy would only have reduced the predicted value to 2.55. It seems reasonable to attribute Hinsz's result to a correlation among the members' recall and interpretation of the video-presented material. For example, a correlation of 0.25 (and optimal weights) would have dropped the predicted performance to the level he observed. One would assume that even higher correlations—and hence even greater performance decreases—would result from choosing jurors from a population having homogenous ethnic and age characteristics.

Another reason for the low overall efficiencies obtained in Experiment 1 was that time pressure may have encouraged the members to use a Condorcet mode of decision. In fact, the results were well below the level predicted by a simple majority Condorcet rule. We attempted to ameliorate this problem in Experiment 2 by removing time pressure as much as possible and by eliminating any financial incentive to rushing the deliberation process. A Condorcet rule is often used when the pressure of work prevents full deliberation and communication among a group's members. For example, consider a group that must select a person for a position from a large number of job applicants. Because of time pressure, the group first selects a "short" final list of candidates from the large applicant pool. This is done individually and without discussion. Each committee member examines the applicants' folders and makes a *yes* or *no* decision on whether to include that applicant on the short list. For a candidate to be on the list, the candidate must receive a favorable vote from all of the committee members. Although the present study offers no solution to the time pressure problem, Sorkin et al.'s (1998) results suggest that the committee would arrive at a better list of candidates if it used a simple majority rule but asked each member to vote for fewer candidates. Note that the American jury does not fit the Condorcet model even though it might use a unanimous rule. In the jury case, member deliberation and interaction are encouraged rather than prohibited.

In addition to handling the correlation and deliberation problems inherent in Experiment 1, Experiment 2 allowed calculation of the separate efficiencies that describe the group decision process and member detection effort. Surprisingly, the efficiency of the group decision process did not decrease with group size. One would expect that increased size would increase problems of coordination and communication, opportunities for disagreements, and social pressure from member subgroups. Apparently, at least in the computer-aided situation of the experimental task, size did not impose significant difficulties. One could argue that the display of member ratings efficiently conveyed most of the information needed by the members to perform the decision task, and that this would probably not be the case for most "real" group decision tasks. We tend to agree with that statement, but we also believe that the present result argues strongly for providing such efficient decision aids in "real" group decision situations.

Why does member effort decrease with group size? We have mentioned the difficulty in measuring a small drop in the detection effort of an individual participant in a group detection task. Only when one aggregates the grouped effect of such drops does one see a clear effect. Individual drops in effort are difficult to observe even when, as in the present experiment, much larger numbers of trials per condition are run than in the typical group study. Because of this statistical aspect, it is relatively easy for an individual in a group to decrease detection effort and remain anonymous. In fact, the group payoff is not affected very much by a decrease in one member's effort, but it is very much affected by a decrease in all the members' efforts.

Kerr (1983), Shepperd (1993), and others discussed why participants may reduce their individual efforts in a group situation and indulge in "social loafing." If some type of social loafing were present, we would conclude that the monetary payoff to each participant was not sufficient to completely dominate incentives to reduce individual effort or participation. Perhaps participants will "loaf" if they cannot see the statistical benefit of their contribution to the group's performance (i.e., to their own payoff) and if they can do so anonymously. Harkins and Petty (1982) found that participants who viewed their contributions as nonessential reduced their individual efforts more than those who viewed their efforts as significant. Following that reasoning, groups with high intermember correlation should show greater decreases in efficiency with size, because the benefit of each member's contribution is very much reduced. We did not run a high correlation condition in Experiment 2, so we cannot test that possibility.

The efficiency of member detection effort (and total group efficiency) shown in Figure 11 appeared to reach an asymptotic level by group sizes of 12 or 14. This result is consistent with social impact theory, which predicts a decreasing rate of reduced effort with increasing group size (Latané, 1991). One is tempted to speculate on the implication and potential application of this result. Suppose that one assumes a single-member detection/decision task having a signal-to-noise ratio of 0.5. A "real-life" version of such a task might involve a predictive decision about the efficacy of a medical research program or the strategic intention of an industrial or military opponent. However, the low signal-to-noise ratio would make this task unfeasible to be performed either by individual observers or small groups of observers. That is, the resulting hit rates would be uselessly low and the false-alarm rates intolerably high. However, if the group efficiency were asymptotic at 0.2 (see Figure 11), one might be able to use the statistical advantage of a large networked group of decision makers to perform the task. For example, a group of 25 decision makers would produce a group d' of $\sqrt{\eta} \cdot \sqrt{m} \cdot \text{SNR} = \sqrt{0.2} \cdot \sqrt{25} \cdot 0.5 = 1.12$. Although not earth-shaking, the resulting performance could be useful in many situations.

Our analysis of group signal detection has emphasized performance accuracy and has avoided the question of individual and group response bias. To specify a system's detection ability completely, it would be necessary to know the ROC, the locus of possible operating points in the hit and false-alarm probability space for the system. This is because the ROC defines the system behavior that will be produced at different levels of bias toward the signal-plus-noise and noise-alone alternatives, including the point of neutral bias. In the present study, our experimental materials and conditions were designed to encourage individuals and groups to operate only at the neutral bias point on the ROC. Our intent was

to avoid the potentially complex problem of aggregating the choices made by detectors having different response biases. To ensure that participants operated with a neutral bias, all detection conditions were conducted with symmetric payoff structures and equal probability of signal plus noise and noise alone. In addition, our participants were knowledgeable about the signal probability and payoff structure, were highly experienced with the experimental conditions, and did not have to record binary responses during the group deliberations. In our experiments they behaved with essentially neutral biases as individuals and as groups.

Minimizing (and controlling) the bias problem allows for a greatly simplified analysis of detection behavior (cf. Pete et al., 1993b; Sorkin & Dai, 1994) but reduces the ability to generalize the results to real-life situations. Moreover, if one assumes that group members are not efficient at finding the optimal aggregate operating point, the neutral-bias condition probably overestimates the performance to be expected from a group that is solving a real-life problem. In other words, having to compensate for variability in member bias provides an additional opportunity for a group to operate in a nonideal manner. Thus, our approach fails to address whether performance is affected by group members being required to form binary judgments before deliberation or how the group sets its criterion when the payoff structure or prior probabilities are not symmetric. We hope to address these questions in future experiments.

References

Ashby, F. G., & Maddox, T. W. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review, 90,* 34–46.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Decision research* (Vol. 2, pp. 103–112). Greenwich, CT: JAI Press.

Berg, B. G. (1989). Analysis of weights in multiple observation tasks. *Journal of the Acoustical Society of America, 86,* 1743–1745.

Berg, B. G. (1990). Observer efficiency and weights in a multiple observation task. *Journal of the Acoustical Society of America, 88,* 149–158.

Berg, B. G., & Green, D. M. (1990). Spectral weights in profile listening. *Journal of the Acoustical Society of America, 88,* 758–766.

Clement, D. E., & Schiereck, J. J. (1973). Sex composition and group performance in a visual signal detection task. *Memory & Cognition, 1,* 251–255.

Condorcet, J. C. (1972). Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix [Essay on the application of analysis to the probability of majority]. New York: Chelsea Publishing Company. (Original work published 1785)

Davis, J. H. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950–1990. *Organizational Behavior and Human Decision Processes, 52,* 3–38.

Durlach, N. I., Braida, L. D., & Ito, Y. (1986). Towards a model for discrimination of broadband signals. *Journal of the Acoustical Society of America, 80,* 60–72.

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

Elvers, G. C., & Sorkin, R. D. (1989). Detection and recognition of multiple visual signals in noise. *Proceedings of the Human Factors Society, 2,* 1383–1387.

Erev, I., Gopher, D., Itkin, R., & Greenshpan, Y. (1995). Toward a generalization of signal detection theory to n-person games: The example of two-person safety problem. *Journal of Mathematical Psychology, 39,* 360–375.

Green, D. M. (1992). The number of components in profile analysis tasks. *Journal of the Acoustical Society of America, 91,* 1616–1623.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grofman, B., Feld, S. L., & Owen, G. (1984). Group size and the performance of a composite group majority: Statistical truths and empirical results. *Organizational Behavior and Human Performance, 33,* 350–359.

Grofman, B., Owen, G., & Feld, S. (1983). Thirteen theorems in search of the truth. *Theory and Decision, 15,* 261–278.

Harkins, S. G., & Petty, R. E. (1982). Effects of task difficulty and task uniqueness on social loafing. *Journal of Personality and Social Psychology, 43,* 1214–1229.

Harvey, N., & Harries, C. (1999, August). Long-term contextual interference in aggregation of opinions: Why does judges' own forecasting impair their use of advice? Paper presented at the Biennial Conference on Subjective Probability, Utility and Decision Making, Mannheim, Germany.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Decision research* (Vol. 2, pp. 129–157). Greenwich, CT: JAI Press.

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion confidence judgments. *Organizational Behavior and Human Decision Processes, 56,* 11–27.

Hillman, B. J., Hessel, S. J., Swensson, R. G., & Herman, P. G. (1977). Improving diagnostic accuracy: A comparison of interactive and Delphi consultations. *Investigative Radiology, 12,* 112–115.

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology, 59,* 705–718.

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121,* 43–64.

Jeffress, L. A., & Robinson, D. E. (1962). Formulas for the coefficient of interaural correlation for noise. *Journal of the Acoustical Society of America, 34,* 1658–1659.

Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology, 45,* 819–828.

Kriss, M., Kinchla, R. A., & Darley, J. M. (1977). A mathematical model for social influences on perceptual judgments. *Journal of Experimental Social Psychology, 13,* 403–420.

Latané, B. (1991). The psychology of social impact. *American Psychologist, 36,* 343–356.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822–832.

Libby, R., Trotman, K. T., & Zimmer, I. (1987). Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology, 72,* 81–87.

Lutfi, R. A. (1995). Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. *Journal of the Acoustical Society of America, 97,* 1333–1334.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge, England: Cambridge University Press.

Metz, C. E., & Shen, J. (1992). Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. *Medical Decision Making, 12,* 60–75.

Montgomery, D. A., & Sorkin, R. D. (1993). The effects of display code and its relation to the optimal decision statistic in visual signal detection.

Proceedings of the Human Factors and Ergonomics Society, 2, 1325–1329.

Montgomery, D. A., & Sorkin, R. D. (1996). Observer sensitivity to element reliability in a multi-element visual display. *Human Factors, 38,* 484–494.

Myung, I. J., Ramamoorti, S., & Bailey, A. D., Jr. (1996). Maximum entropy aggregation of expert predictions. *Management Science, 42,* 1420–1436.

Nitzan, S., & Paroush, J. (1982). Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review, 23,* 289–297.

Nitzan, S., & Paroush, J. (1984). A general theorem and eight corollaries in search of a correct decision. *Theory and Decision, 17,* 211–220.

Pete, A., Pattipati, K. R., & Kleinman, D. L. (1993a). Distributed detection in teams with partial information: A normative-descriptive model. *IEEE Transactions on Systems, Man, and Cybernetics, 23,* 1626–1647.

Pete, A., Pattipati, K. R., & Kleinman, D. L. (1993b). Optimal team and individual decision rules in uncertain dichotomous situations. *Public Choice, 75,* 205–230.

Richards, V. M., & Zhu, S. (1994). Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *Journal of the Acoustical Society of America, 95,* 423–434.

Robinson, D. E., & Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. In R. Eberts & C. G. Eberts (Eds.), *Trends in Ergonomics/Human Factors* (Vol. II, pp. 75–82). Amsterdam: North-Holland.

Shapley, L., & Grofman, B. (1984). Optimizing group judgmental accuracy. *Public Choice, 43,* 329–343.

Shepperd, J. A. (1993). Productivity loss in performance groups: A motivation analysis. *Psychological Bulletin, 113,* 67–81.

Sorkin, R. D. (1990). Perception of temporal patterns defined by tonal sequences. *Journal of the Acoustical Society of America, 87,* 1695–1701.

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Journal of Organizational Behavior and Human Decision Processes, 60,* 1–13.

Sorkin, R. D., Mabry, T. R., Weldon, M., & Elvers, G. (1991). Integration of information from multiple element display. *Organizational Behavior and Human Decision Processes, 49,* 167–187.

Sorkin, R. D., Robinson, D. E., & Berg, B. G. (1987). A detection theory method for evaluating visual and auditory displays. *Proceedings of the Human Factors Society, 2,* 1184–1188.

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science, 9,* 456–463.

Swets, J. A. (1984). Mathematical models of attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 183–242). Orlando, FL: Academic Press.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnosis. Collected papers.* Hillsdale, NJ: Erlbaum.

Tanner, W. P., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *Journal of the Acoustical Society of America, 30,* 922–928.

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

Appendix

Performance of a Group That Uses Weights $\{a\}$

The performance, d', of a system that uses the Z statistic and arbitrary weights, $\{a_i\}$, is given by the expected value (over trials) of Z given signal minus the expected value of Z given noise, all divided by the standard deviation of Z. Let $X_{i,j}$ be the value of detector i's estimate on the jth trial. The expected value of Z given noise is 0, so the numerator for d' is just the expected value of Z given signal. That is,

$$E[Z|_{\text{signal}}] - E[Z|_{\text{noise}}] = E[\sum_{i=1}^{m} a_i(X_{i,j}|_{\text{signal}})] = \sum_{i=1}^{m} a_i d_i'$$

The variance of Z over trials is the same given signal plus noise or noise, and because the independent and common components of X are independent,

$$\begin{aligned}\text{Var}[Z] &= \text{Var}[\sum_i a_i X_{i,j}] = \text{Var}[\sum_i a_i X_{\text{IND},j} + \sum_i a_i X_{\text{COM},j}] \\ &= \text{Var}[\sum_i a_i X_{\text{IND},j}] + \text{Var}[\sum_i a_i X_{\text{COM},j}] \\ &= \sum_i [\text{Var}(a_i X_{\text{IND},j})] + \text{Var}[(\sum_i a_i) X_{\text{COM},j}] \\ &= \sum_i [a_i^2 \text{ Var}(X_{\text{IND},j})] + (\sum_i a_i)^2 \text{ Var}[X_{\text{COM},j}] \\ &= \sigma_{\text{IND}}^2 \sum_i a_i^2 + \sigma_{\text{COM}}^2 (\sum_i a_i)^2 \\ &= (1-\rho) \sum_i a_i^2 + \rho(\sum_i a_i)^2\end{aligned}$$

and,

$$d'_{\text{weight}} = \frac{\sum_{i=1}^{m} a_i d_i'}{\sqrt{(1-\rho) \sum_{i=1}^{m} a_i^2 + \rho(\sum_{i=1}^{m} a_i)^2}}.$$

Received November 30, 1995
Revision received August 6, 1999
Accepted June 15, 2000 ■