ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 33, 350–359 (1984)

Group Size and the Performance of a Composite Group Majority: Statistical Truths and Empirical Results

BERNARD GROFMAN

University of California, Irvine

SCOTT L. FELD

State University of New York, Stony Brook

AND

GUILLERMO OWEN

University of California, Irvine

The judgmental accuracy of group majority decision making in binary choice, multiple-item prediction tasks as a function of group size, average competence of group members, and the shape of the overall distribution of judgmental accuracy in the group are discussed. For the case where the addition of an $(N + 1)$th member to a group of size N has a known cost function and the value of a correct group decision can be specified, the optimal group size is calculated.

I. INTRODUCTION

In our paper we consider conditions under which additional group members can help or hurt group accuracy. In a recent article in this journal, Libby and Blashfield (1978) report results of an empirical study of the effects of group size upon the accuracy of judgments/predictions based upon decisions of a "hypothetical composite" group majority or upon a "hypothetical composite" group modal choice. The results indicated that "on average the majority of the increment gained by aggregating large numbers of judges can be obtained by aggregating three judges," and the authors go on to say that "this suggests that in situations where expected performance was the only criterion, it would be unlikely that employment of larger aggregates would be cost-beneficial" (1978, p. 278). They also assert that "in cases where the individual judges are not very accurate, the increment in accuracy gained by increasing composite size declines more rapidly than in cases where judges are more accurate"

Requests for reprints should be sent to Dr. Bernard Grofman, School of Social Sciences, University of California, Irvine, CA 92717.

0030-5037/84 $3.00

(1978, p. 128). Libby and Blashfield do not discuss in any detail statistical models which could account for their results, although they do review a number of earlier studies on the superiority of group to individual judgment.

We present new results on group judgmental accuracy in single- and multiple-item prediction tasks. The model we offer allows us to specify the convergence of judgmental accuracy to a limiting value as a function of group size, which was observed by Libby and Blashfield for various multiple-item tasks. Moreover, our model allows us to derive the exact value of that asymptote and the rate of convergence as a function of the mean ability of group members and the distribution of individual judgmental-predictive competence. These results lead to some modification of the Libby and Blashfield claims. Furthermore, for the case where the addition of an (N + 1)th individual to a group of size N has a known cost function and the value of a correct group decision can be specified, we show how to calculate the optimal group size.

The basic theorem on the effect of pooling of judgments is due to the French mathematician and philosopher, the Marquis de Condorcet (1785). This theorem, which can be thought of as a variation of the well-known "law of large numbers," was, however, "lost" for a number of years until rediscovered by Black (1958), although it is now well known. (For a history of the theorem, see Grofman, 1975; for extensions see Grofman, Owen, & Feld, 1982a, 1982b; and Owen, Grofman, & Feld, 1982).

In order to report results concisely, we introduce the following notation: For any dichotomous choice task, average group competence shall be denoted $\bar{p}$ ($0 \leq \bar{p} \leq 1$); the competence of individual group members is denoted p_i, and the accuracy of a majority vote in a group of size N is P_n. For simplicity, unless otherwise stated, we assume N to be odd. The Condorcet theorem in its original form assumes that group members are homogeneous; i.e., $p_i = p_j = p$ for all i, j. When all group members are identical in judgmental competence, we drop the subscripts.

Condorcet Jury Theorem

If $1 > p > 1/2$, then P_N is monotonically increasing in N and $\lim_{N\to\infty} P_N \to 1$; *if* $0 < p < 1/2$, then P_N is monotonically decreasing in N and $\lim_{N\to\infty} P_N \to 0$; if $p = 1/2$, then $P_N = 1/2$ for all N (for a proof see Black, 1958, or Grofman, 1978). If $p > 1/2$, this theorem can be interpreted as "*vox populi, vox dei.*" It is rather remarkable how fast P_N goes up (down) with N if $p \neq 1/2$. We show results for $N = 1$ through 19 in Table 1, from Grofman (1975). Owen et al. (1982) have generalized this result for any distribution of p_i to show that we can replace p with $\bar{p}$ in the statement of the theorem above.

TABLE 1
THE PROBABILITY THAT A GROUP MAJORITY WILL REACH A CORRECT JUDGMENT FOR VARIOUS VALUES OF N AND p[a]

	p				
N	.2	.4	.5	.6	.8
1	.2000	.4000	.5000	.6000	.8000
3	.1040	.3520	.5000	.6480	.8960
5	.0580	.3174	.5000	.6826	.9420
7	.0335	.2858	.5000	.7102	.9666
9	.0196	.2666	.5000	.7334	.9804
11	.0116	.2466	.5000	.7534	.9884
13	.0070	.2288	.5000	.7712	.9930
15	.0042	.2132	.5000	.7868	.9958
17	.0026	.1990	.5000	.8010	.9974
19	.0016	.1860	.5000	.8140	.9984

Source. Grofman, 1975.

[a] N = group size, p = the probability that an individual member of the group will reach a correct judgment in a dichotomous choice situation.

II. APPLICATIONS OF THE CONDORCET JURY THEOREM

Libby and Blashfield (1978, p. 128) assert that "In cases where the individual judges are not very accurate, the increment in accuracy gained by increasing composite size declines more rapidly than in cases where judges are more accurate." Actually, even for the simple binary choice case, the situation is somewhat more complex. We show values of ΔP_N ($= P_N - P_{N-2}$) for $p = .6$ and $p = .8$ in Table 2.

TABLE 2
ΔP_N THE INCREMENT IN THE PROBABILITY THAT A GROUP MAJORITY WILL REACH A CORRECT JUDGMENT, AS A FUNCTION OF p AND N

	p	
N	.6	.8
1	.6000	.8000
3	.0480	.0960
5	.0346	.0460
7	.0276	.0246
9	.0232	.0138
11	.0200	.0080
13	.0178	.0046
15	.0156	.0028
17	.0142	.0013
19	.0130	.0010

As we can see from Table 2, while the increment in judgmental accuracy with increasing group size is *initially* larger for $p = .8$, *once* $N = 7$ is reached, the increase in judgmental accuracy obtained by adding additional members to the group diminishes rapidly for $p = .8$ and slowly for $p = .6$. We can precisely specify the nature of the incremental process by obtaining an iterative expression for P_{N+2} as a simple function of P_N, p, and N.

Result 1 (Recursion Formula for Condorcet Jury Theorem)[1]

$$\Delta P_{N+2} = P_{N+2} - P_N \cong \frac{2^{N+2}(p - 1/2)\, e - 2N(p - 1/2)^2}{\sqrt{\pi N}} \qquad (1)$$

Thus, as asserted in Libby and Blashfield (1978, p. 123), "Reliability of a composite will increase as a negatively accelerated exponential function of the number of judges." We also see from Eq. (1) that the closer individual competence is to 1/2, the slower will be the rise in group competence with increasing N. The maximum will occur where the positive effect of $2^{N+2}(p - 1/2)$ is equal to the negative effect of $e^{-2N(p-1/2)^2}$. This will, in general, occur for small values of N.

III. MULTI-ITEM DECISION TASKS

Let us now look at what happens when individuals (and statistically created groups) are asked to deal with multi-item decision tasks. Our earlier results can be applied, but we must proceed with some caution in distinguishing between aggregate performance and the performance of aggregates. Some additional notation will also be helpful.

Consider a (perhaps statistically created) group of N individuals on a multi-item test or prediction task, T, involving a series of dichotomous choice items. Let us use P_{N_T} to denote the proportion of items that a hypothetical group majority or test taker gets correct on that exam. Let p_{i_T} denote the proportion of correctly answered questions by the ith individual on test T (we need not expect that all individuals will score the same). Let $\bar{p}_T$ denote the average proportion correct (over people) on test T.

It might appear that we can simply substitute $\bar{p}_T$ for $\bar{p}$ (or P) in our earlier expressions, and P_{N_T} for P_N. This would, however, be erroneous except where all items are the same in difficulty. More generally, for any values of N, it is quite possible for $\bar{p}_T > 1/2$ and yet $P_{N_T} < 1/2$ or for $\bar{p}_T < 1/2$ and yet $P_{N_T} > 1/2$ because any multi-item task may be thought of as consisting of *difficult* and *easy* items. For a given subject pool, difficult

[1] For proof of Result 1 see Appendix.

TABLE 3

		Items ($j = 1,2, \ldots, J$)				
		1	2	3		
People ($i = 1,2, \ldots, N$)	1	1	0	0	$\overline{P}_1 = .33$	
	2	1	0	0	$\overline{P}_2 = .33$	$P_{N_T} = .33$
	3	1	1	1	$\overline{P}_3 = 1$	
Ease =		1	$\frac{1}{3}$	$\frac{1}{3}$	$\overline{P}_T = .56$	

items are simply those which more than half the test takers get wrong, and easy items are those which more than half the test takers get right. Even if there are more difficult questions than easy questions, if the difficulty of items is not extreme while the easy items are all relatively easy, then the average score over individuals can be greater than .5 and yet most questions are answered wrong most of the time. (Henceforth, we shall omit the T subscript whenever it is clear from the context that a multi-item test is meant.) Consider, for example, a three-item test where $\overline{p}_1 = \overline{p}_2 = .33$ and $\overline{p}_3 = 1$ (Table 3). Here $\overline{p} = .56$; yet two of the three questions can be expected to be answered wrongly by a group using a majority rule strategy. In like manner, even if there are more easy questions than difficult questions, if the difficult items are very difficult while the easy items are not too easy, then the average score $\overline{p}$ can be less than .5 and yet most questions are answered correctly by a majority (cf. Grofman, 1981b).

Let us denote the fraction of hard (difficult) questions on a multi-item test (decision/prediction task) as H and use E to denote the fraction of easy questions.

Result 2

For a multi-item test consisting of dichotomous choices in which there are no questions at .5 difficulty, then

$$\lim_{N \to \infty} P_N = E. \tag{2}$$

Proof. The proof follows straightforwardly from the Condorcet jury theorem as extended by Owen *et al.* (1982).

Most standardized exams have a substantial portion of items which are more apt to be answered wrongly than answered correctly (see, e.g., Lord, 1980, Fig. 2.2.1, 13). Most multi-item classification or prediction tasks also have items which vary (often greatly) in difficulty. (This is true, for example, for the data base in Libby, 1976.) Thus, we would expect

that the performance of a (statistically created) group majority would go asymptotically toward a value E, considerably less than 1. Indeed, if $H > E$, then performance would *decrease* with increasing group size.

The rate of convergence of the performance of the group majority to its ceiling E will vary as a function of the distribution of item difficulty. It is straightforward to demonstrate the general result that the rate of convergence is rapid when all items are either very hard and/or very easy and is slow when some or all items have difficulty level near to .5. Also, by symmetry, the impact of an extra person on rapidity of convergence on an item with $p = .8$ will be the same as for an item with $p = .2$, only in the opposite direction. In particular, if the distribution of item difficulty is perfectly symmetric around .5, then adding additional test takers does nothing to affect accuracy. If there is a normal distribution of test difficulty items, the rate of convergence will be monotonic. The same obtains for any unimodal symmetric distribution.

For the case where the cost of adding an additional "judge" can be specified and where the value of a correct decision is known, our method of analysis permits us a straightforward expression for the calculation of optimal group size.

Result 3 (Optimum Group Size)

Let the value of a correct decision be denoted V and the cost of utilizing an additional group member be c. For simplicity we shall assume c is not a function of N, but this simplification is not critical.[2] The optimal group size is simply the maximum value of N such that

$$(P_N - P_{N-2})\, V > 2c. \tag{3}$$

It is instructive to work out a simple case of the application of Eq. (3). Let us consider a 100-item true–false exam, where (for a specified subject pool) 60% of the questions will on average be answered correctly by 80% of the people and 40% of the questions will on average be answered correctly by 40% of the people. Here $E = .6$. For this case, we show in Table 4 values of ΔP_N for $N = 1,3,5, \ldots, 11$. We have deliberately chosen an example which is nonmonotonic in P_N because of an asymmetry in the rates of convergence on hard and easy items. In this example, it would *never* be desirable to use more than 7 judges. If $V =$ \$1000, $c =$ \$10, the optimum group size is 3, since (.0146) (1000) < 20. If $V =$ \$1000, $c =$ \$2, the optimum group size is still only 5.

Of course if we maintain $E = .6$ and $H = .4$ but change $\bar{p}_H$ from .4 to .2, then (because of symmetry) we will obtain P_N monotonic in N. For

[2] We shall neglect asymmetry in Type I and Type II errors, although it is reasonably straightforward to cope with such complications (see, e.g., Grofman, 1979, 1981a).

TABLE 4

P_N AND ΔP_N VALUES FOR AN EXAM WITH 60% EASY QUESTIONS WITH $\bar{p}_E = .8$ AND 40% HARD QUESTIONS WITH $\bar{p}_H = .4$; AND AN EXAM WITH 60% EASY QUESTIONS WITH $\bar{p}_E = .8$ AND 40% HARD QUESTIONS WITH $\bar{p}_H = .2$

N	P_N	ΔP_N $(=P_N - P_{N-2})$	N	P_N	ΔP_N
1	.6400	.6400	1	.5600	.5600
3	.6776	.0376	3	.5792	.0192
5	.6922	.0146	5	.5884	.0092
7	.6953	.0031	7	.5934	.0050
9	.6949	−.0004	9	.5961	.0027
11	.6917	−.0032	11	.5977	.0016
13	.6873	−.0044	13	.5986	.0011
15	.6828	−.0045	15	.5992	.0006
17	.6780	−.0048	17	.5995	.0003
19	.6734	−.0046	19	.5997	.0002
	$\bar{p}_E = .8, \bar{p}_H = .4$			$\bar{p}_E = .8, \bar{p}_H = .2$	

that case we have results as shown in the second part of Table 4. *In general, it need not be true that only a handful of judges is optimal.* If the average difficulty level is near 1/2, then the rate of convergence to the asymptote may be quite slow. (See Table 1.) If V is very large relative to c, the marginal gain from additional members may be considerable. For example, if $V = \$1000$ and $c = \$0.50$, then the optimum group size for the second example shown in Table 4 is $N = 13$. Hence, it is misleading to claim, as did Libby and Blashfield (1978), that $N = 3$ will in general be optimum.

In general, nonmonotonicity in P_N is to be expected. However, as is apparent from Table 4, differences in P_N for neighboring values of N can be so small as to make nonmonotonicity virtually impossible to detect—since its presence will be masked by stochastic noise.

IV. DISCUSSION

We have shown how, for dichotomous choices, the accuracy of a (statisticized) group majority can be estimated as a function of group size and of the proportions of easy and difficult questions with which the group is confronted—where easy and difficult are used in a special technical sense defined above. The methods we have used can be extended to choice among more than two alternatives in a reasonably straightforward fashion by treating polychotomous choice as a sequence of pairwise decisions (see Farquharson, 1969). Our results can be used to explain recent empirical results on the relationship between group size and judg-

mental accuracy, such as the work by Libby and Blashfield (1978).[3] More importantly, they show how empirical results may really be statistical tautologies in disguise. Understanding these tautologies can lead to specifying the conditions required for various results to hold so that valid generalizations can be made.

APPENDIX: PROOF OF RESULT 1

After the addition of two members to a group of size N, the only way for P_{N+2} and P_N to differ is if the new members have changed the direction of the group majority. This can only occur when previously the group's decision was reached by exactly a majority *and* the two new voters are in agreement. The probability that a bare group majority was wrong is

$$\binom{N}{\frac{N+1}{2}} p^{\left(\frac{N-1}{2}\right)} (1-p)^{\left(\frac{N+1}{2}\right)};$$

the probability that a bare group majority was right is

$$\binom{N}{\frac{N+1}{2}} p^{\left(\frac{N+1}{2}\right)} (1-p)^{\left(\frac{N-1}{2}\right)}.$$

[3] Rather than simply looking at the total number of correct predictions, various authors have used alternative measures. We believe it most sensible to judge accuracy in terms of the proportion of correct predictions for dichotomous choice decisions where marginals are equal; but for unequal marginals or where there are multiple alternatives to choose from, and perhaps also "degrees" of correctness which can be assigned to the various possible answers, it is useful to look at other measures (cf. Libby, 1976). A measure of group and individual predictive accuracy used by Libby and Blashfield (1978) is the phi coefficient. In one task they examined, 20 individuals with subject area expertise were asked to evaluate 99 applicants based upon a 44-variable data base including demographic and academic information and to predict the screening decisions of an admission committee charged with responsibility for PhD program admissions. In another task, 43 individuals with relevant experience postdicted whether 60 business firms would experience failure within 3 years. Exactly half of these firms actually had gone out of existence. Information presented included various standard financial data about the firms. Libby and Blashfield (1978) present results of median phi values for majority group decisions in groups of various sizes for the business firm failure postdiction task and for the graduate admission task. For dichotomous decisions the phi coefficient is identical to Pearson's r. For the former task, for $N = 1$ phi was .4842; for $N = 3$ phi was .5670; while for $N = 43$ phi has risen to only .6365. For the latter task, for $N = 1$ phi was .4598, for $N = 3$ phi was .5239, while for $N = 20$ phi has risen to only .5953. Without the raw data we cannot know the actual p_j values for these two tasks. We can, however, show exactly how the phi coefficients should vary with E and N and then estimate E from the observed asymptotic value of phi. If we take the asymptotic values of phi to be .64 and .60, respectively, in the two decision tasks examined by Libby and Blashfield (1978), we obtain the asymptotic values of P_N as .82 and .80 for the two cases. For these cases we obtain estimated E values of .82 and .80, respectively, since the asymptotic value of the phi coefficient is simply $2E-1$.

The probability that both new group members are right is p^2; the probability that both are wrong is $(1-p)^2$. Hence, from Bayes theorem on conditional probabilities, we have

$$\Delta P_{N+2} = P_{N+2} - P_N = p^2\binom{N}{\frac{N+1}{2}}p^{\left(\frac{n-1}{2}\right)}(1-p)^{\left(\frac{N+1}{2}\right)} \tag{1A}$$

$$-(1-p)^2\binom{N}{\frac{N+1}{2}}p^{\left(\frac{N+1}{2}\right)}(1-p)^{\left(\frac{N-1}{2}\right)},$$

$$= (2p-1)\binom{N}{\frac{N+1}{2}}p^{\left(\frac{N+1}{2}\right)}(1-p)^{\left(\frac{N-1}{2}\right)}. \tag{2A}$$

Using Stirling's formulas, we have

$$\Delta P_{N+2} \cong (2p-1)\left[\frac{2}{\sqrt{\pi N}}p^{\left(\frac{N+1}{2}\right)}(1-p)^{\left(\frac{N-1}{2}\right)}\right]. \tag{3A}$$

If we let $p = (1 + 2q/2)$ and $1 - p = (1 - 2q/2)$ we may reexpress Eq. (3A) as

$$\Delta P_{N+2} \cong \frac{2^{N+1}(2p-1)(1-4q^2)^{N/2}}{\pi N}. \tag{4A}$$

From the Taylor expansion of the natural logarithm we have

$$\log(1-4q^2) = -4q^2 + 8q^4 - \frac{64}{3}q^6 + \ldots .$$

Substituting the first term of this expansion in Eq. (4A) and then taking antilogs and substituting back $q = p - 1/2$, we obtain the desired new approximation

$$\Delta P_{N+2} = \frac{2^{N+2}(p-1/2)\,e^{-2N(p-1/2)^2}}{\sqrt{\pi N}}. \tag{5A}$$

REFERENCES

Black, D. (1958). *The theory of committees and elections*. London: Cambridge Univ. Press.

Condorcet, N. C. de. (1785). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix*. Paris.

Farquharson, R. (1969). *Theory of voting*. New Haven, CN: Yale Univ. Press.

Grofman, B. (1975). A comment on 'democratic theory: A preliminary mathematical model.' *Public Choice*, 21, 100–103.

Grofman, B. (1978). Judgmental competence of individuals and groups in a dichotomous choice situation. *Journal of Mathematical Sociology,* **6,** 1, 47–60.

Grofman, B. (1979), A preliminary model of jury decision making as a function of jury size, effective jury decision rule, and mean juror judgmental competence. In G. Tullock (Ed.), *Frontiers of economics.* Blacksburg, VA: Center for Study of Public Choice, Virginia Polytechnic Institute and State University.

Grofman, B. (1981a). Mathematical models of jury/juror decision making. In Bruce D. Sales (Ed.), *Perspectives in law and psychology,* Vol. II: *The jury, judicial and trial processes (pp. 305–351). New York: Plenum.*

Grofman, B. (1981b). Can most of the voters be outvoted most of the time? Unpublished manuscript.

Grofman, B., Owen, G., & Feld, S. L. (1982a). Thirteen theorems in search of the truth. *Theory and Decision.*

Grofman, B., Owen, G., & Feld, S. L. (1982b). Average competence, variability in competence and the accuracy of statistically pooled decisions. *Psychological Reports,* **50,** 683–688.

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance,* **16,** 13–22.

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

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Owen, G., Grofman, B., & Feld, S. L. (1982). *A theorem on the optimal distribution of competence within a group.* Unpublished manuscript.

RECEIVED: March 1, 1982.