PSYCHOMETRIKA—VOL. 20, NO. 2
JUNE, 1955

TWO MODELS OF GROUP BEHAVIOR IN THE SOLUTION OF EUREKA-TYPE PROBLEMS*

IRVING LORGE
AND
HERBERT SOLOMON
TEACHERS COLLEGE, COLUMBIA UNIVERSITY

A study by Shaw (7) some twenty years ago is frequently cited by social scientists to support the generalization that groups are superior to individuals in problem-solving. Shaw suggests that personal interaction within the group is responsible for the superior performance of groups. This article re-examines her data in the light of two models which propose that the difference in quality of solution between group and individual performance is solely a matter of ability. It is shown that Shaw's data may be considered to have been an outcome of behavior postulated by the models. Since Shaw's observations relate to a special population and to special kinds of problems, the proposed models may not be appropriate under differing experimental conditions. In fact, Lorge *et al.* (4) have indicated that experimental demonstration of the superiority of groups over individuals in problem-solving depends not only on the kind of group but also on the kind of problem to be solved. In addition, the diversity of transfer of training for groups and for individuals is considered.

Introduction

Since this article treats only the data from the first half of the Shaw experiments, a brief description of this part will be given. Three problems (**3**), each a well-known mathematical puzzle involving the transport of objects under certain constraints, were given to groups and to individuals. The first, known historically as the Tartaglia, requires the transport of three jealous husbands and their three beautiful wives across a river in a boat holding just three at a time, under the constraint that no husband will allow his wife in the presence of another man unless he is also present, and with the specification that only husbands can row. The second problem, the historical Alcuin, is similar in that it requires the transport of three missionaries and three cannibals in a boat carrying two at a time under the constraint that missionaries may never be outnumbered by cannibals, and with the specification that all missionaries and just one cannibal have mastered the art of rowing. The third problem, the historical Tower of Hanoi, or disc problem, is similar to the previous two in that it requires the transport of three graduated discs, stacked in order of size, to another position via an intermediate way station, under the constraint that a larger disc may never

*Supported in part by the Office of Naval Research under Contract N6 onr 266 (21) and the Air Force Personnel and Training Research Center under Contract AF 18(600)-341.

be placed on a smaller one, with the specification that only one disc may be moved at a time.

Shaw's subjects were students in a social psychology class which had been divided into halves: one half being formed at random into *ad hoc* like-sex, four-member groups, and the other half serving as individuals, i.e., as controls. Thus, the performances of five groups were contrasted with those of twenty-one individuals. Each group and each individual was asked to solve all three problems in the same sequence.

A criterion for comparing group and individual performance is the contrast between the proportion of individuals and the proportion of groups successful in the solution of each problem. For Shaw's three problems, the proportions of individuals and groups mastering each solution are given in Table 1 (Columns 1 and 3). When, for each problem separately, the difference between proportions of success in groups and in individuals is tested, using an upper one-sided .05 critical region, the data for Problems I and II support the generalization of group superiority, but the difference between groups and individuals for Problem III is not statistically significant. The statistical test (**2**, **6**) of the hypothesis that two proportions are equal is

$$z = \frac{\theta_G - \theta_I}{\sqrt{\dfrac{1}{N_I} + \dfrac{1}{N_G}}}, \tag{1}$$

where $\theta = 2 \arcsin \sqrt{p}$, p = proportion of success, N = sample size, and the subscripts I and G refer to individuals and to groups, respectively. The function z is approximately normally distributed with zero mean and unit variance under the hypothesis tested. The results of this analysis could be used to support Shaw's conclusion (**7**, p. 504): "Groups seem assured of a much larger proportion of correct solutions than individuals do."

Of the five groups, however, two solve *none* of the problems and two solve *all* problems. Of the twenty-one individuals, none solves more than *one* of the three problems. The fact that some groups solved none and some groups solved all the problems suggests the hypothesis that the observed group superiority is due to the abilities of the members of the group rather than personal interaction. Such an hypothesis may be expressed in terms of two ability models: (A) group superiority is a function only of the ability of one or more of its members to solve the problem without taking account of the interpersonal rejection and acceptance of suggestions among its members; (B) group superiority is a function only of the pooled abilities of its members. The latter model, B, implies that any problem may be composed of, and solved in, two or more stages. Model B, of course, reduces to Model A for one-stage problems.

Model A

Under Model A the probability of a group solution is the probability that the group contains one or more members who can solve the problem. This non-interactional ability model for any specific problem can be expressed mathematically as follows: Let

P_G = the probability that a group of size k solve the problem;
P_I = the probability that an individual solve the problem.

Then

$$P_G = 1 - (1 - P_I)^k, \tag{2}$$

where P_G and P_I are population parameters considered fixed for the specific problem and the specific population.

Confidence in the tenability of this non-interactional ability model can be decided by testing it on the basis of sample observations. Assume N_G observations of group performance and N_I of individual performance. Then sample estimates p_G and p_I may be obtained, where p_G and p_I are the ratios of the observed successes to attempts for groups and for individuals, respectively; p_G should be compared with p_{G_A} (or equivalently, p_I with p_{I_A}), where

$$p_{G_A} = 1 - (1 - p_I)^k, \tag{3}$$

or equivalently

$$p_{I_A} = 1 - (1 - p_G)^{1/k}. \tag{3a}$$

The observed difference $(p_G - p_{G_A})$ certainly can be used as a test of the model, for the smaller the observed difference, the more tenable is the model and, the larger the observed difference, the less tenable it is. If an α level of significance is used, then the model would be rejected if

$$\Pr \{(p_G - p_{G_A}) > O_d\} \leqq \alpha$$

and accepted otherwise, where O_d is the observed difference. A one-sided test is used since *negative* personal interaction (an unable majority preventing an able minority from solving the problem) is not anticipated in the Shaw groups, and thus the test is made most powerful against all alternatives indicating *positive* personal interaction. That is, if positive interaction does exist, the probability of rejecting Model A is higher than the probability given by a two-sided test of the same size. A similar argument holds for $(p_{I_A} - p_I)$, since it is an equivalent test.

To test the existence of the model, the distribution of $(p_G - p_{G_A})$ must be obtained. Although p_G and p_{G_A} are independently distributed proportions, the distribution of their difference is no longer related to the standard distribution of the difference of two binomials since p_{G_A} is not a binomial; p_{G_A} is a function of p_I, which is a binomial. This complicates obtaining the exact distribution of $(p_G - p_{G_A})$ either in closed form or in a form such that existing tables may be used. Since sample sizes are small, however, it is not too tedious to compute the exact probabilities of all differences larger than the observed difference under the assumptions that (1) the model holds and (2) the nuisance parameter (either P_G or P_I) is replaced by a sample estimate.

It is interesting to note that

$$E(p_{G_A}) = 1 - (1 - P_I)^4 - \frac{6P_I(1 - P_I)^3}{N_I} + \frac{P_I(1 - P_I)^2(4 - 11P_I)}{N_I^2} - \frac{P_I(1 - P_I)(1 - 6P_I + 6P_I^2)}{N_I^3},$$

and

$$\sigma^2_{p_{G_A}} = \frac{P_I(1 - P_I)}{N_I}[16(1 - P_I)^6] + \frac{f_1(P_I)}{N_I^2} + \cdots + \frac{f_6(P_I)}{N_I^7},$$

where $f_j(P_I)$, $j = 1, 2, \cdots, 6$, are eighth-degree polynomials in P_I. Thus, for large N_I, p_{G_A} is an unbiased estimate of P_G and its variance is $16(1 - P_I)^6$ $\sigma^2_{p_I}$.

For the three Shaw problems, there are six possible values for p_G and twenty-two possible values of p_I. In Problem I, for instance, the observed difference $(p_G - p_{G_A})$ is .14, where p_{G_A} is computed from formula (2) using the value of p_I reported by Shaw. It is necessary, therefore, to tabulate all possible differences greater than the value .14. For these tabulated differences, the probability of each is computed under the specified assumptions. The probability for each difference is the product of the probabilities that the p_G and p_{G_A} involved in the difference do occur when the two assumptions hold. The probability that a p_{G_A} occurs is equal to the probability that its corresponding p_I occurs. The probability for p_G and p_{G_A} may be obtained readily by reference to a binomial table (5). The sum of these products of probabilities is the exact probability that an observed difference will exceed .14. In Table 1, column five gives the exact probability, P, that the observed difference $(p_G - p_{G_A})$ will be exceeded by chance.

An approximation to the exact probability can be made when p_I is small enough so that p_{G_A} can be approximated by kp_I, for then

$$\frac{1}{k}(2 \arcsin \sqrt{kp_I}) \quad \text{and} \quad \frac{1}{k}(2 \arcsin \sqrt{p_G})$$

are approximately normally distributed with variances

$$\frac{1}{N_I} \quad \text{and} \quad \frac{1}{k^2 N_G}, \text{ respectively.}$$

Thus, if Model A holds,

$$z = \frac{2 \arcsin \sqrt{p_G} - 2 \arcsin \sqrt{kp_I}}{\sqrt{\frac{1}{N_G} + \frac{k^2}{N_I}}} \tag{4}$$

is approximately normally distributed with zero mean and unit variance. Some liberties have been taken in this approximation by assuming kp_I to be binomial since it can assume values greater than one. This assumption, apparently, does not impair its usefulness for the Shaw experiments. In Table 1, column six gives $P' = P_r\{z > z_0\}$, where z_0 is the specific value for z corresponding to the observed difference. Notice that the approximation obviously gets better as p_I decreases.

The hypothesized non-interactional ability Model A, thus, is rejected for Problem II, but accepted as tenable for Problems I and III. For each of the three problems, however, p_G exceeds p_{G_A}, suggesting that Model A might be modified and improved.

TABLE 1

	p_I	p_{I_A}	p_G	p_{G_A}	P	P'
Problem I	3/21 = .14	.20	3/5 = .60	.46	.38	.48
Problem II	0/21 = .00	.20	3/5 = .60	.00	.029	.023
Problem III	2/21 = .095	.12	2/5 = .40	.33	.43	.48

p_I = ratio of individual solutions to attempts
p_G = ratio of group solutions to attempts
p_{I_A} = estimate of P_I from Model A and observation p_G
p_{G_A} = estimate of P_G from Model A and observation p_I
P = probability $(p_G - p_{G_A})$ is exceeded by chance under Model A and P_G or P_I is replaced by sample estimate
P' = approximation of P replacing p_{G_A} by kp_I

Stage-wise Solutions

Within the framework of strict ability models, a modification of Model A may be made. Solution of eureka-type problems may be considered the consequence of pooling success at each of several stages of the problem. Shaw's study, indeed, suggests the plausibility of such a stage-wise model. In reporting about the erroneous moves made by her subjects in solving Problem I she states that 13 different individuals made an error in the first move, four made an error in the third move, and one made an error in the fifth. For groups, however, she reports "No group erred on the first move; one erred on the third and one on the fourth."

Shaw's description of the errors in Problem I suggests the importance of the first move, since 13 of the 21 individuals failed to make the correct first move. Each group, however, apparently had in it at least one member who made the first move successfully since none of the five groups erred on it. Once the first move is accomplished, the difficulty of the problem changes. Five individuals who made the first move correctly did fail at subsequent stages, i.e., made the first correct move but failed at later moves. Two groups failed at some later move, suggesting that the group lacked at least one member who could accomplish some later move.

Assuming that a problem is solved in s independent stages, (not the moves Shaw mentions, since such moves may be interrelated) and assuming that Model A (equation 2) applies at each stage j, then,

$$P_G = \prod_{j=1}^{s} [1 - (1 - P_{I_j})^k], \qquad P_I = \prod_{j=1}^{s} P_{I_j} , \tag{5}$$

where s is the number of stages, and P_{I_j} is the probability of success for an individual at stage j. Now for the purpose of estimating s from the Shaw data, consider the assumption that P_{I_j} is the same for each stage; thus $P_{I_j} = P_I^{1/s}$, then

$$P_G = [1 - (1 - P_I^{1/s})^k]^s. \tag{5a}$$

This assumption may possibly be unrealistic, but it is necessary to provide an estimate of s from Shaw's data.

Substituting the estimates for P_G and P_I from Shaw's Problems I and III, $s = 2$ (to the nearest integer) for both problems; for Problem I, $s = 1.6$; for Problem III, $s = 1.5$. Since the observed proportions of individual solutions for Problem II is zero, s is indeterminate. (If for Problem II, P_I is replaced by $p_{I_A} = .2$, then s is very close to 1.)

It is not too difficult to rationalize the two-stage nature of the problems. For example, in the problem of the jealous husbands and their wives, the basic first stage requires the recognition that the boat, which may carry three, must be limited to taking just a husband and his wife across the river. Once this first stage is solved, the second and final stage is analogous to repetitious knitting. It is interesting to note that if it is assumed that $p_I = .05$ and $p_G = .95$ (an indication of overwhelming group superiority through positive personal interaction among its members) then by (5a) $s = 10$ to the nearest integer, an estimate even larger than the number of moves required in some of the Shaw problems. While all possible pairs of the values, p_G and p_I have not been considered, an excessively large difference gives a value of s inconsistent with a psychological analysis of the problem into steps or stages.

Model B

On both a probabilistic and a content basis, a two-stage problem may be reasonably inferred; assume now that Problems I, II, III are two-stage problems. For this situation, the population of individuals may be classified in the following way:

Population Type	*Ability*	*Proportion in the Population*
X_1	Solve both stages	P_1
X_2	Solve stage 1, not stage 2	P_2
X_3	Solve stage 2, not stage 1	P_3
X_4	Solve neither stage	P_4

Assuming this multinomial distribution of ability, appropriate ability interaction within a group of four individuals can accomplish a solution even though the group has in it no one member who can solve the problem as a whole; for example, the group whose members symbolically are represented as $X_2\ X_3\ X_3\ X_3$. Consider all possible samples of four $(X_i\ X_j\ X_k\ X_m)$ from this population. It is possible to enumerate all groups of four that can interact to accomplish whole solutions solely by pooling their abilities. Any group containing at least individual X_1, or at least individuals X_2 and X_3 jointly, will be successful. The probability of occurrence of each sample of four is given by the multinomial distribution if P_1, P_2, P_3, P_4 are known. The sum of probabilities of the occurrence of each group of four that can complete a stage-wise solution is the probability of a group solution on the hypothesis of stage-wise pooling of ability. Thus, under Model *B*, the probability of a group solution is obtained by a special summation of the elements of the multinomial distribution.

Currently, not enough knowledge is available for estimating all the probabilities P_1, P_2, P_3, and P_4. At best, in line with current knowledge of the distribution of ability, the psychologist can merely supply reasonable estimates for P_2, P_3, and P_4. In Shaw's data, P_1 can be estimated from the sample. This still leaves two degrees of freedom for choices since the sum of the four probabilities is one.

Suppose these two free choices are subject to the restriction that they closely reproduce p_G and that they are not inconsistent with psychological knowledge of the distribution of ability. For the kind of problems treated by Shaw, psychological evidence indicates that the percentage of persons who will fail on both stages will be larger than the percentage who can solve both stages or any one stage. This, of course, does not uniquely determine the four parameters but it is interesting to see that reasonable estimates do exist. For example, if in Shaw's Problem I, $P_I = .15$ ($p_I = .1428$), $P_2 = .15$, $P_3 = .15$, and $P_4 = .55$, then $p_{G_B} = .61$ as contrasted with $p_G = .60$. Here P_2, P_3, and P_4 were guesses to reproduce the observed p_G. They are also not inconsistent with the distribution of ability. Actually p_G can be reproduced exactly, but it was not considered necessary to alter the p_i's slightly to accomplish this since enough leeway has already been taken to reproduce a *sample* value. Moreover, slight changes would not alter any decisions about the reasonableness of the P_i's. This argument also applies in the following discussion of Problems II and III. Incidentally, $P_2 = P_3 = .15$ leads to

$P_1 + P_3 = P_2 + P_3 = .30$; this indicates that the probability of an individual's success in stage 1 and stage 2 is .30. By (5a), $P_G = .58$ as contrasted with $p_G = .60$, which suggests that the assumption which yields (5a) from (5) is realistic after all.

Moreover, if in Shaw's Problem I, $P_1 = .15$, $P_2 = .30$, $P_3 = .30$, and $P_4 = .25$, a situation definitely inconsistent with the distribution of ability, we get $p_{G_B} = .92$, a value noticeably different from $p_G = .60$.

Similarly for Problem III, if $P_1 = .10$ ($p_1 = .0952$), $P_2 = P_3 = .10$, $P_4 = .70$, then $p_{G_B} = .42$, as contrasted with $p_G = .40$. Also referring to (5a), $P_G = .35$, as contrasted with $p_G = .40$. In Problem II, $P_1 = .2$, $P_2 = P_3 = .05$, and $P_4 = .70$ yields $p_{G_B} = .61$, as contrasted with $p_{G_B} = .60$; again referring to (5a), $P_G = .46$, as contrasted with $p_G = .60$. It should be noticed here that this big difference arises from the use of $p_{I_A} = .2$, which would lead to a one-stage problem if it were p_I. Notice that this is reflected also in $P_2 = P_3 = .05$, for $P_2 = P_3 = 0$ is a one-stage model. Substitituion of the unrealistic observed $p_I = 0$ would yield nonsensical results. This information is presented in Table 2.

It is interesting to note the premium gained by the two-stage model. Model B can be made to account for most of the excess $(p_G - p_{G_A})$ not accounted for by Model A. If Model B holds, the excess is the probability of a group solution when individual X_1 is not in the group of 4. For the weights described, this is .13 for Problem I, and .077 for Problem III; these should be compared with $(p_G - p_{G_A}) = .14$ for Problem I, and .07 for Problem

TABLE 2

	Problems		
	I	II	III
p_G	.60	.60	.40
p_{G_A}	.46	.00	.33
p_{G_B}	.61	.61	.42
P_1	.15	.20	.10
P_2	.15	.05	.10
P_3	.15	.05	.10
P_4	.55	.70	.70

p_I = ratio of individual solutions to attempts
p_G = ratio of group solutions to attempts
p_{G_A} = estimate of P_G from Model A and observation p_I
p_{G_B} = estimate of P_G from Model B and weights P_1, P_2, P_3, and P_4

P_1 = probability that an individual will solve both stages in Model B
P_2 = probability that an individual will solve stage 1 but not stage 2 in Model B
P_3 = probability that an individual will solve stage 2 but not stage 1 in Model B
P_4 = probability that an individual will not solve either stage in Model B

III. For Problem II, the weights used lead to an excess of .017, but this is just another reflection of the fact that the replacement of p_I by p_{I_A} leads to a one-stage problem.

The stage-wise model hypothesizing the pooling of ability tends to reproduce the observed p_G when reasonable weights are used. Indeed, unreasonable weights produce major discrepancies from the observed p_G . The implication of the model is that group superiority may be conceived as a function only of pooling the abilities of its members. Ultimately, empirical estimates must be obtained for P_2 , P_3 , and P_4 . One experimental procedure for such estimates would require individuals to solve the problem. For instance, in a two-stage problem, those individuals solving the problem (as in the Shaw data) provide a basis for estimating P_1 . Some individuals who failed the whole problem, however, will have accomplished stage 1 successively but failed on stage 2, providing a basis for estimating P_2 . The remainder, those who could not accomplish stage 1, would be given the problem reduced by the accomplishment of stage 1, reported as a fact, with the requirement that the "new" problem be solved. Some of the individuals will then solve the "new" problem providing a basis for estimating P_3 .

When P_1 , P_2 , P_3 , and P_4 are estimated by p_1 , p_2 , p_3 , and p_4 on the basis of sample observations, assuming Model B holds, a value p_{G_B} will be obtained and contrasted with p_G . As in Model A, the probability that an observed difference will be exceeded by chance must be computed in order to examine the tenability of the model. Under the assumption that Model B holds, and replacing P_1 , P_2 , P_3 , and P_4 by their estimates, it is possible to obtain the exact distribution of $(p_G - p_{G_B})$, although it is extremely tedious to compute. If p_i is based on n_i observations, then p_{G_B} can assume $(n_1 + 1) \cdot (n_2 + 1) \cdot (n_3 + 1) \cdot (n_4 + 1)$ values. Even if the sample sizes are small, say $n_i = 5$, p_{G_B} takes on 1296 values. This, plus the difficulty of actually computing the probability of a difference $(p_{G_B} - p_G)$, renders the technique somewhat useless. Moreover, for large samples an asymptotic method seems fruitless because of the special way the multinomial distribution is summed for this situation.

Suppose, however, a confidence interval for P_G , say P_{G_L} and P_{G_U} is obtained from p_G . Assuming the model holds, all the sets p_1 , p_2 , p_3 , p_4 which yield values between P_{G_L} and P_{G_U} inclusive, form a confidence region for P_1 , P_2 , P_3 , P_4 . Actually all that need be done then is to consider the value p_{G_B} yielded by the observed p_1 , p_2 , p_3 , p_4 . If this value lies between P_{G_L} and P_{G_U} the model is tenable for the specified confidence coefficient employed, let us say $1 - \alpha$, or equivalently for the significance level α.

Pooling of Data

Shaw pools the results for the three problems, neglecting the fact that the *same* individuals and *same* groups worked the three problems in the same sequence. Thus, she contrasts 8/15 or 53 per cent success for groups with 5/63 or 7.9 per cent success for individuals. Using the z test given by (1) with the awareness that the lack of independence renders it inadequate, this difference is statistically significant at the 5 per cent level. Moreover, since the correlation between observations can be assumed to be positive, the decision of statistical significance is on the conservative side. Also, Model A is rejected using the z test given by (4). It should be emphasized that of the five groups, two solve none of the three problems and two solve all. Of the twenty-one individuals, none solves more than one of the three problems! Two alternate hypotheses are suggested: 1) Model B is operating; 2) groups do better than individuals in a sequential solution of problems of the same kind. Hypothesis 2 can arise from three possibilities: (a) negative transfer in individuals, zero or positive transfer in groups; (b) zero transfer in individuals, positive transfer in groups; (c) positive transfer in individuals, greater positive transfer in groups. As regards hypothesis 2, Cook (1), using two versions of the disc problem (Problem III), varying in difficulty of sequence, implies "that transfer 'spuriously' lowers the probability of a given individual achieving the same degree of success or failure (relative to the rest of the groups) on both problems." The evidence from Shaw's groups suggests somewhat the same conclusion by indicating the plausibility of positive transfer in groups in sequential solution of problems of the same kind. A carefully designed experiment to ascertain the superiority of groups over individuals in transfer of training is suggested by this combined evidence.

REFERENCES

1. Cook, T. W. Amount of material and difficulty of problem solving. II. The disc transfer problem. *J. exp. Psychol.*, 1937, **20**, 288-296.
2. Eisenhart, C. Inverse sine transformation of proportions. In Eisenhart, Hastay, and Wallis, Techniques of statistical analysis, New York: McGraw-Hill, 1947.
3. Kasner, E. and Newman, J. Mathematics and the imagination. New York: Simon and Schuster, 1940.
4. Lorge, I., Fox, D., Davitz, J., Weltz, P., Herrold, K. Products of individual and of group: Review of literature. Unpublished manuscript developed in connection with Contract AF 33 (038) 28792, Human Resources Research Institute, Maxwell Air Force Base, Alabama.
5. National Bureau of Standards. Tables of the binomial probability distribution, Applied Mathematics Series 6, January 1950.
6. Paulson, E. and Wallis, A. Planning and analyzing experiments for comparing two percentages. In Eisenhart, Hastay, and Wallis, Techniques of statistical analysis, McGraw-Hill, 1947.
7. Shaw, M. E. Comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, **44**, 491-504.

Manuscript received 5/17/54

Revised manuscript received 9/16/54