9-1-1968

Regularity in the Distribution of Participation in Group Discussions

Joseph B. Kadane

Gordon H. Lewis

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut

COWLES FOUNDATION DISCUSSION PAPER NO. 255

REGULARITY IN THE DISTRIBUTION OF PARTICIPATION IN GROUP DISCUSSIONS

Joseph B. Kadane and Gordon H. Lewis

September 23, 1968

ABSTRACT

Reports have been published indicating striking regularities in the distribution of participation in group discussions, regularities potentially important in the understanding of small group processes. Several attempts to explain the form of the regularity contain implications for many sociological analyses. These attempts to explain the regularity are found to be inadequate; at the same time, the regularity is found to be somewhat doubtful.

REGULARITY IN THE DISTRIBUTION OF PARTICIPATION IN GROUP DISCUSSIONS

by

Joseph B. Kadane and Gordon H. Lewis*

In the early fifties reports were published of regularity in the distribution of participation in group discussions (Bales et al., 1951; Stephan and Mishler, 1952). It was first suggested that the regularity might be approximated by a harmonic function, but this notion was finally rejected (Bales et al., 1951). Subsequently, it was suggested that the regularity could be approximated by an exponential function (Stephan and Mishler, 1952; Stephan, 1952). Regularities are important to science because they can lead to theories which explain the regularities and which also imply additional hypotheses of regularities. It is these twin aspects of explanation and prediction which lead to a cumulation of knowledge.

Since publication of the regularities, there have been three attempts to explain why the regularity appears to have an exponential form. Two of these (Coleman, 1960; Leik, 1967) attempt to explain the regularity by using statements which are on the same level as the statement of the regularity, statements about the entire group meeting. In contrast, the third (Horvath, 1965) attempts to explain the regularity by focusing not on the

* This research was partially supported by a grant from the National Science Foundation.

entire group meeting but upon the individual units of which the entire group discussion consists. Since Horvath's model is discussed elsewhere (Kadane et al., 1968), the present paper is confined to those models having the entire meeting of a group as the unit of analysis.

The quantitative regularity reported by Stephan and Mishler appears to be a substantive law. If Coleman is correct, the regularity is a by-product of the way the data were handled; if Leik is correct, the regularity is due to characteristics of individuals comprising the group and the population from which the persons were drawn. In neither case is the regularity considered as representing processes of interaction in the group. We find neither explanation adequate to account for the alleged regularity, but we also find the regularity itself to be in doubt.

Background

The groups studied by Bales et al. (1951) included students working on contrived problems, therapy groups, and "case-discussion meetings of diagnostic councils operating in a research setting" (1951:461). During the meetings trained observers recorded each person's participation, the unit of participation being the "simple subject-predicate combination," if the act were verbal, or the "smallest overt segment of behavior that has 'meaning' to others in the group" if the act were nonverbal (1951:462). Following the observation of a number of groups, Bales et al. ranked the members of each group according to the number of acts they had initiated. These ordered sets of frequencies were aggregated for goups of the same size and presented graphically as percentage distributions.

Although Bales et al. were disinclined to believe that the harmonic function would fit their data, they included the estimates, $X_{(i,n)}$, from the harmonic function:

$$X_{(i,n)} = \frac{S}{i \sum_{j=1}^{n} \frac{1}{j}} \qquad (1)$$

where

$X_{(i,n)}$ is the expected number of acts by the person with the i^{th} rank in a group of n persons, and

S is the sum of acts in the set of data under consideration (1951:467).

To evaluate the fit of the data to the harmonic function Bales et al. presented the observed and expected frequency distributions for groups of size six and an evaluative criterion based on a chi square measure.[1] They concluded that for groups of size six "the fit is not sufficiently good to permit us to believe that the deviations have arisen as random fluctuation" (1951:467); the same conclusion was reached for the other group sizes. In their remarks about the harmonic function they suggest that it is obviously too much to expect that a function which depends only on the size of the group should accurately portray the regularities in the distribution of participation.

The groups studied by Stephan and Mishler were considerably more homogeneous; they were classes at Princeton, each class consisting of an instructor and a set of students. During the class meeting observers recorded each person's participation, in this case the unit of participation being

> the word, sentence, or longer statement of an individual that follows such a participation by one member and continues until it is terminated by an appreciable pause or by the participation of another member (1952:600).

The procedure followed by Stephan and Mishler in aggregating their data was the same as that used by Bales et al. The frequencies of participation were ordered for the members of a group, the ordered distributions were aggregated across groups of the same size, and the final distributions were converted to percentages. These distributions are produced in Table 1. In every case the highest participation is by the set of instructors, or "leaders."

Table 1

Exponential Approximation to Stephan and Mishler's Data

Persons of Rank	Size of Group, Including Leader: 4		5		6		7		8	
	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.
L	42.8		44.4		42.7		45.6		39.3	
1	28.1	29.5	25.0	25.1	23.9	24.7	20.8	20.1	21.3	21.3
2	20.5	17.4	14.7	15.4	15.4	14.9	12.3	13.3	14.4	14.3
3	8.6	10.3	11.7	9.4	10.1	9.0	8.8	8.8	9.0	9.7
4			4.2	5.8	5.9	5.4	6.0	5.8	6.9	6.4
5					2.0	3.3	4.3	3.8	4.8	4.3
6							2.3	2.5	2.7	2.9
7									1.7	1.9
Participations	755		856		2951		1999		2042	
Sessions	6		7		17		15		14	
Mean Error	2.07		1.14		.84		.44		.30	
$\hat{r}$	.5907		.6125		.6036		.6614		.6702	

Persons of Rank	Size of Group, Including Leader: 9		10		11		12	
	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.
L	45.0		49.1		51.3		48.8	
1	18.8	19.1	17.3	16.2	14.2	14.6	14.5	14.4
2	12.4	12.7	10.7	11.2	9.8	10.4	10.5	10.5
3	8.9	8.5	7.5	7.8	8.2	7.4	8.0	7.6
4	6.2	5.7	6.4	5.4	5.5	5.2	6.3	5.5
5	4.4	3.8	3.1	3.8	3.9	3.7	4.8	4.0
6	2.8	2.5	2.7	2.6	2.5	2.6	3.2	2.9
7	1.0	1.7	1.9	1.8	2.3	1.9	1.7	2.1
8	0.6	1.1	1.5	1.3	1.1	1.3	1.7	1.6
9			0.0	0.9	0.7	0.9	0.6	1.1
10					0.5	0.7	0.0	0.8
11							0.0	0.6
Participations	1269		481		437		525	
Sessions	10		4		3		5	
Mean Error	.47		.55		.34		.44	
$\hat{r}$	.6664		.6948		.7102		.7275	

It is obvious in Table 1 that as the group size increases there is a growing disparity between the percentage for the leaders and the percentage for the highest ranking set of students. But, Stephan and Mishler suggested, if one omits the percentage contributed by the set of leaders, p_L, the remaining portion of the distribution is well approximated by the exponential function

$$p_i = ar^{i-1} \qquad (2)$$

where p_i is the estimated percentage for the <u>students</u> ranked i, a and r are parameters depending on the size of the group, and $\Sigma p_i = 100 - p_L$. Mishler justified the exclusion of the data from the instructors on several grounds:

> First, (the instructors) differed from the students in knowledge of the subjects under discussion, experience in discussion, etc. Second, they had different functions to perform in the meeting, a different role to play. Third, it was deemed a sufficient first step to find a function that fitted the student members' participation rates, without the addition of another function for the leaders' roles (1952:602).

One problem, of course, is to find estimates of the parameters a and r. Stephan and Mishler rejected the use of maximum likelihood estimates of these parameters since it was "not possible to establish a defensible probability model" (1952:602).[2] Wanting

> to give what was judged to be appropriate weight to the fit for large and for small percentages, the function was fitted to the data by minimizing the sum of squares of deviations of the logarithms of the estimated percentages from the logarithms of the observed percentages, each square being weighted by the observed percentage. That is, the quantity to be minimized was:
>
> $$\sum_{i=1}^{n} p_i \, [\log p_i - \log (ar^{i-1})]^2$$
>
> where p_i is the percentage observed for the (students) ranked i (1952:602).

By taking the partial derivatives of this function with respect to a and to r and setting them equal to zero, Stephan and Mishler obtained estimates for a and for r, viz.:

$$\log \hat{a} = \frac{CD - BE}{AC - B^2}$$

$$\log \hat{r} = \frac{AE - BD}{AC - B^2} \qquad (3)$$

where $A = \Sigma p_i$, $B = \Sigma i p_i$, $C = \Sigma i^2 p_i$, $D = \Sigma p_i \log p_i$, $E = \Sigma i p_i \log p_i$.

From equation (2), with separate estimates $\hat{a}$ and $\hat{r}$ for each group size, Stephan and Mishler calculated the percentage distribution, $\{\hat{p}_i\}$, for the student members of the group. The estimated participation of the leader was obtained residually,

$$\hat{p}_L = 100 - \sum_{i=1}^{n} \hat{p}_i \ .$$

It is not clear, however, why separate estimates were desired for a and r. If r (or $\hat{r}$) is known, then for any subset of values of the type ar^k , a is determined. For example, let

$$ar^i + ar^j + \ldots + ar^m = C \ ,$$

then

$$a = \frac{C}{r^i + r^j + \ldots + r^m} \ .$$

A special case of this arises when $i-1$ is the exponent of the ith term. Using the fact that

$$\sum_{i=0}^{n-1} X^i = \frac{1-X^n}{1-X}$$

one obtains

$$a = \frac{1-r}{1-r^n} \ C \qquad (4)$$

In the present case, $C = 100 - p_L$.[3] Conversely, given a (or $\hat{a}$) there is a unique value for r.

The observed percentage distributions in Table 1 have been recalculated from the frequency distribution published by Stephan and Mishler; the estimated distributions in Table 1 are based on the exponential function with $\hat{r}$ given by (3) and $\hat{a}$ by (4).[4] Since the discrepancy between the observed and estimated percentages for the leader is automatically zero, the estimated value for the leader has been omitted from Table 1. For the other members of the group, however, the absolute differences between the observed and estimated distributions have been summed and divided by the number of student members to give some idea of the fit of the exponential function. The mean error ranges from .30% to 2.07% with a mean across groups of .74% (cf. Table 1). These errors were apparently small enough to lead Stephan and Mishler to conclude that the exponential fit the data adequately well.

In 1952 Stephan published another paper in which he suggested that the exponential function fit not only the data he and Mishler had gathered, but the data of Bales et al. as well. Although apparently the only data for which he was able to make a direct comparison were those for groups of size six, Stephan concluded that the function fit "quite well" (1952:485).

To facilitate a fuller comparison of the fit of the exponential function, Table 2 contains both the observed and the estimated percentage distributions for Bales' data.[5] The parameter $\hat{r}$ is again determined by the Stephan and Mishler procedure, but in this case it is based on data for members of all ranks in the group, and the estimated values, $\hat{p}_i$, are generated for members of all ranks, none being predicted residually. The fit of the function is clearly better for the smaller groups than for the larger ones, the reverse of what was found for the Stephan and Mishler data. The mean errors range from .33% to 3.95% with an average across groups of 2.33%. This is more than three times the average error when the exponential function was applied to the Stephan and Mishler data (.74%).

Since it has been alleged that the exponential function approximates both data sets, it is important to consider why it seems to give so much better fit for Stephan and Mishler's data than for Bales'. Here two differences clearly related to the fit are examined: the inclusion of data for the highest participators and the degree of homogeneity of the groups.

First, if one omits the act initiated by the set of highest participators in the Bales' groups, the mean error across groups drops to .58% (cf. Table 3). The only mean error greater than .68% is that for group size nine, and that distribution is based

Table 2

Exponential Approximation to Bales' Data

	Size of Group							
Persons of Rank	3		4		5		6	
	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.
1	44.38	44.63	32.90	33.32	46.12	43.42	43.02	37.62
2	32.67	32.18	27.27	27.01	22.04	26.03	18.72	24.56
3	22.95	23.20	22.68	21.90	15.63	15.60	14.20	16.04
4			17.14	17.75	10.46	9.35	11.04	10.47
5					5.75	5.60	7.43	6.84
6							5.59	4.47
Acts	9,304		58,218		10,714		21,311	
Sessions	26		89		9		18	
Mean Error	.33		.52		1.59		2.56	
$\hat{r}$	.7210		.8107		.5994		.6530	

	Size of Group							
Persons of Rank	7		8		9		10	
	Obs.	Est.	Obs.	Est.	Obs.	Est.	Obs.	Est.
1	43.14	34.02	39.78	33.01	49.08	35.65	42.61	29.29
2	15.24	23.25	16.56	22.65	18.99	23.21	11.94	21.02
3	11.90	15.89	12.70	15.54	7.59	15.11	10.06	15.09
4	9.88	10.86	9.84	10.66	5.34	9.83	9.00	10.83
5	8.57	7.42	8.62	7.31	4.92	6.40	6.20	7.77
6	6.30	5.07	5.48	5.02	4.08	4.17	5.28	5.58
7	4.96	3.47	4.26	3.44	3.80	2.71	5.06	4.00
8			2.74	2.36	3.66	1.77	3.68	2.87
9					2.53	1.45	3.29	2.06
10							2.87	1.48
Acts	22,044		12,830		1,422		2,823	
Sessions	15		10		1		3	
Mean Error	3.71		2.43		3.95		3.56	
$\hat{r}$	.6835		.6860		.6510		.7176	

Table 3

Exponential Approximation to Bales' Data: Highest Participator Excluded

Persons of Rank	3 Obs.	3 Est.	4 Obs.	4 Est.	5 Obs.	5 Est.	6 Obs.	6 Est.
	Size of Group							
1	44.38		32.90		46.12		43.02	
2	32.67	32.67	27.27	27.62	22.04	22.70	18.72	19.03
3	22.95	22.95	22.68	21.98	15.63	14.92	14.20	14.09
4			17.14	17.49	10.46	9.81	11.04	10.43
5					5.75	6.45	7.43	7.72
6							5.59	5.71
Acts	9,304		58,218		10,714		21,311	
Sessions	26		89		9		18	
Mean Error	0.0*		.47		.68		.29	
$\hat{r}$	.7025		.7958		.6574		.7402	

Persons of Rank	7 Obs.	7 Est.	8 Obs.	8 Est.	9 Obs.	9 Est.	10 Obs.	10 Est.
	Size of Group							
1	43.14		39.78		49.08		42.61	
2	15.24	15.18	16.56	16.95	18.99	14.16	11.94	12.02
3	11.90	12.24	12.70	12.87	7.59	10.61	10.06	9.97
4	9.88	9.87	9.84	9.78	5.34	7.96	9.00	8.27
5	8.57	7.96	8.62	7.42	4.92	5.97	6.20	6.86
6	6.30	6.42	5.48	5.64	4.08	4.47	5.28	5.69
7	4.96	5.17	4.26	4.28	3.80	3.36	5.06	4.72
8			2.74	3.25	3.66	2.52	3.68	3.92
9					2.53	1.89	3.29	3.25
10							2.87	2.69
Acts	22,044		12,830		1,422		2,823	
Sessions	15		10		1		3	
Mean Error	.23		.36		1.77		.31	
$\hat{r}$	.8063		.7594		.7499		.8295	

*Zero by definition.

on the observation of a single session. Thus, omitting the acts initiated by the highest ranking participants clearly gives results more consistent with the exponential function.[6] But before concluding that the exponential holds only when the highest rank participants have been omitted, it is well to remember that in Table 2 the exponential function fits quite well for groups of size 3 and 4 even when the highest participators are present.

Second, in some cases aggregating separate data sets obscures regularities. An example of this is contained in Table 4. These data come from a recent study by Bonacich (1968).[7] The table includes the fit of the exponential for groups which were composed of males, groups composed of females, and the combined groups. The mean error (1.16%) for the combined groups is slightly lower than that for groups of size five in Table 2 (1.59%), but there is a very big difference in the mean error for males compared to the mean error for females (.41% vs. 1.57%). The value for males seems quite comparable to the best fit for any data set of Bales or of Stephan and Mishler and is only one third the mean error for male and female groups combined. The poorer fit for Bales' data, therefore, may have resulted from the aggregation of heterogeneous groups.[8]

Before closing this discussion of the exponential relation it might be well to include data from three additional studies which also show an equally good fit to the exponential, cf. Table 5.

TABLE 4

Exponential Approximation to Bonacich's Data

Groups of Size 5

Persons of Rank	Males Obs.	Males Est.	Females Obs.	Females Est.	Combined Obs.	Combined Est.
1	38.56	38.97	33.87	35.74	35.40	36.79
2	25.55	25.66	26.38	25.19	26.11	25.36
3	17.93	16.90	19.64	17.75	19.08	17.48
4	11.10	11.13	13.35	12.51	12.62	12.05
5	6.85	7.33	6.76	8.82	6.79	8.31
Acts	14,136		29,158		43,294	
Sessions	20		36		56	
Mean Error	.41		1.57		1.16	
$\hat{r}$	.6586		.7048		.6894	

TABLE 5

Exponential Approximation of Participation: Additional Evidence

Persons of Rank	Chicago		Harvard		Yale	
	Obs.	Est.	Obs.	Est.	Obs.	Est.
1	43.79	44.16	36.65	36.93	35.92	36.14
2	33.01	32.27	27.31	27.46	27.44	27.38
3	23.20	23.57	21.63	20.43	21.32	20.75
4			14.42	15.19	15.31	15.72
Acts	39,045		14,875		11,410	
Sessions	67		24		15	
Mean Error	.50		.60		.32	
$\hat{r}$	.7306		.7438		.7576	

The data from three person groups come from a study of family groups (a father, mother, and child);[9] the data from four person groups come from studies of undergraduates at Harvard and at Yale.[10]

The conclusion seems to have been that the exponential function fits the data much better than the harmonic function does. But it is important to notice that in evaluating the fit of the harmonic function Bales et al. applied a chi square criterion to the observed and estimated frequencies, in evaluating the fit of the exponential, Stephan and Mishler used the absolute difference between the observed and estimated percentages. It is interesting to consider what decision might have been reached had the fit of the two functions been evaluated using a single criterion, and of the two criteria, the absolute error seems preferable because the chi square test requires the assumption of independence.

Table 6 contains the observed values for Bales' data and the estimates based on the harmonic function (1) for percentage distributions, i.e. where S = 100. Using the mean percentage error across groups, the harmonic does almost as well as the exponential. The mean error across groups for the exponential was 2.33% and for the harmonic it is 3.13%. But if one considers all groups larger than size four, the harmonic does better than the exponential (1.79% versus 2.97%), and in particular, when the

TABLE 6

Harmonic Approximation to Bales' Data

Persons of Rank	3 Obs.	3 Est.	4 Obs.	4 Est.	5 Obs.	5 Est.	6 Obs.	6 Est.
1	44.38	54.56	32.90	48.01	46.12	43.80	43.02	40.82
2	32.67	27.28	27.27	24.00	22.04	21.90	18.72	20.41
3	22.95	18.19	22.68	16.00	15.63	14.60	14.20	13.61
4			17.14	12.00	10.46	10.95	11.04	10.20
5					5.75	8.76	7.43	8.16
6							5.59	6.80
Mean Error	6.76		7.55		1.36		1.20	

Persons of Rank	7 Obs.	7 Est.	8 Obs.	8 Est.	9 Obs.	9 Est.	10 Obs.	10 Est.
1	43.14	38.57	39.78	36.79	49.08	35.35	42.61	34.14
2	15.24	19.28	16.56	18.40	18.99	17.67	11.94	17.07
3	11.90	12.86	12.70	12.26	7.59	11.78	10.06	11.38
4	9.88	9.64	9.84	9.20	5.34	8.84	9.00	8.54
5	8.57	7.71	8.62	7.36	4.92	7.07	6.20	6.83
6	6.30	6.43	5.48	6.13	4.08	5.89	5.28	5.69
7	4.96	5.51	4.26	5.25	3.80	5.05	5.06	4.88
8			2.74	4.59	3.66	4.42	3.68	4.27
9					2.53	3.93	3.29	3.79
10							2.87	3.41
Mean Error	1.64		1.33		3.35		1.85	

entire distribution is estimated for groups of size six, the groups extensively analyzed by Stephan (1952), the harmonic does better than the exponential (1.20% versus 2.56%). On the other hand, when applied to the student members in the groups of Stephan and Mishler, the exponential does better than the harmonic. Across all groups sizes the mean error for the exponential is .74% and for the harmonic 1.66%. None of these differences is compelling evidence in favor of the exponential.

Even if the fit of harmonic and the exponential functions is judged to be identical, however, there would appear to be reason to believe that one should adopt the exponential function rather than the harmonic. Only the exponential function has been placed in an explanatory schema, and it would seem ill advised to reject a function for which there exists an explanation in favor of one for which there does not. Thus the acceptance of the exponential function rests on the explanations which have been produced.

Three separate explanations have been given for the exponential nature of the distribution of participation. Horvath (1965) presented a stochastic model of the process of participation from which it can be shown that the exponential relation follows. Results of recent tests (cf. Kadane et al., 1968), however, cast great doubt on the adequacy of that stochastic model.

The other two explanations are based not on a process model of participation but on the methodology of the studies. Coleman (1960:54) has suggested that it is the act of ordering and aggregating the data that may have produced the exponential regularity; Leik (1967:285) has suggested that random sampling from a particular distribution may have produced the regularity. In both cases the attempt is made to explain the exponential regularity as a function of the research methods. Evaluation of these arguments follows.

Coleman's Model

In reviewing the work of Bales and Stephan and Mishler, Coleman paid particular attention to the characteristic of the exponential function that the ratio p_i/p_{i-1} is a constant. If the empirical distributions were truly characterized by exponential relations these values would be constant for a particular size group, which they are not, as Table 7 shows. Nevertheless, it is to this characteristic of the exponential that Coleman referred when he asked

> whether the constant ratio between ranks might not have occurred if random numbers had been aggregated in a similar way. In other words, suppose the participations were distributed randomly among the _n_ members --- might not the regularity have resulted anyway, simply as a consequence of reordering the members by their participation rates and then aggregating? (1960:54).

The model which he proposed is most easily described as an equal probability, n-state, independent trials process model. That is, on any trial the probability of going from state i to state j is $1/n$, for all i and j. The states in this case are participants, and the process being in state i at time k is

TABLE 7

Constancy of the Ratio r

$$r = p_i / p_{i-1}$$

Stephan and Mishler's Data

Size of Group, Excluding Leader

i	3	4	5	6	7	8	9	10	11
2	.731	.589	.646	.589	.677	.660	.614	.694	.724
3	.419	.794	.653	.714	.626	.720	.706	.837	.764
4		.360	.589	.680	.761	.699	.861	.667	.786
5			.337	.723	.700	.709	.484	.708	.758
6				.535	.561	.625	.867	.647	.680
7					.618	.371	.692	.909	.529
8						.538	.778	.500	1.000
9							0	.600	.333
10								.667	0
11									0

Bales' Data

Size of Group

i	3	4	5	6	7	8	9	10
2	.734	.830	.477	.436	.352	.417	.387	.282
3	.706	.832	.709	.755	.783	.759	.400	.833
4		.753	.673	.782	.832	.786	.697	.910
5			.552	.676	.869	.869	.925	.659
6				.747	.733	.640	.837	.883
7					.794	.764	.927	.943
8						.643	.974	.740
9							.676	.892
10								.848

the same as person i making the kth act. The question is, if a given number of participations is distributed in accordance with the model, and if the resulting frequencies are ordered and summed across a number of experiments, would the aggregate resemble the exponential function?

Since the expected values for the entire process were not known, Coleman performed random number experiments, the results of which are in some respects similar to the participation data. The similarities were primarily (1) that "the ratios, $r_{i,i+1}$ (sic) for the statistical data are about as constant as those for the empirical participation data, (2) that "the deviations from the average ratio seem to have a similar pattern to the empirical data: negative deviations for the last one or two ratios," and (3) that "in some cases, the entire pattern of deviations of the statistical ratios (i.e., the shape of the graph) follows closely that of the empirical data: (1960:56).

The major difference, however, was that in every case the average ratio of adjacent frequencies was higher than that for the empirical data, even though the number of acts simulated per group was only 25. Coleman compared this to an average of 70 acts per group (by the students) for the groups by Stephan and Mishler, but the comparison would have been even more striking had he made it to the average of over 800 acts per group in Bales' groups. Coleman's conclusion was that if the process had operated for a larger number of trials, the average ratio for the statistical data would be even more divergent from the average ratio for the empirical data than it is. Thus, if a more realistic number of acts were simulated, the model would indicate greater dissimilarity to the data obtained empirically.

This is not terribly surprising, however, when one reflects on the nature of the model Coleman used. There are two main features of that model: (1) the random process which was used to generate the frequencies, and (2) the ordering and aggregating of these frequencies. The random process Coleman used had the characteristic that as M, the total number of acts simulated, increases the proportion of acts initiated by person i approaches the value $1/n$, where n is the number of persons in the group. Thus the variation of these "random numbers" decreases as the number of acts increases.

A modified Coleman model which limits itself to ordering and aggregating random numbers is of considerable interest. Assume that one takes random samples from a uniform distribution of numbers, orders these numbers, and aggregates them. Would such a model result in an exponential function?

It is important to note that in addition to eliminating Coleman's random process for generating the numbers, the modified Coleman model has no restriction that the number of acts in one group be the same as that in another.[11] This change appears to present a closer correspondence to the operations followed by Bales and by Stephan and Mishler; it gives more weight to groups with larger numbers of acts.

Fortunately, the modified Coleman model can be solved without resort to Monte-Carlo experimentation. Since ratios of expected values of order statistics are unchanged if the underlying distribution is multiplied by any positive constant, the uniform distribution may be taken on the interval $[a, a+1]$. The kth largest of a sample of size n from a uniform distribution is known to have a beta distribution with parameters $n-k+1$ and espectation $a + \frac{n-k+1}{n+1}$, thus, the proportion of participation has probability limit

$$\frac{a + \frac{n-k+1}{n+1}}{\sum_{j=1}^{n} \left(a + \frac{n-j+1}{n+1}\right)} = \frac{a + \frac{n-k+1}{n+1}}{n\left(a + \frac{1}{2}\right)}$$

which is linear in k, not exponential.

Table 8 gives the expected values (as percentages) when the numbers are from the interval [0, 1].[12] For groups larger than size five the percentage expected by the highest initiator is noticeably underestimated.

TABLE 8

Percentage Distributions Produced by the Modified Coleman Model

	n							
i	3	4	5	6	7	8	9	10
1	50.00	40.00	33.33	28.57	25.00	22.22	20.00	18.18
2	33.33	30.00	26.67	23.81	21.43	19.44	17.77	16.36
3	16.67	20.00	20.00	19.05	17.86	16.66	15.55	14.54
4		10.00	13.33	14.29	14.29	13.88	13.33	12.73
5			6.67	9.52	10.71	11.11	11.11	10.91
6				4.76	7.14	8.33	8.89	9.09
7					3.57	5.55	6.67	7.27
8						2.78	4.44	5.45
9							2.22	3.64
10								1.82

The ratio of the kth largest expectation to the kth-1 is

$$\frac{a(n+1)+n-k+1}{a(n+1)+n-k+2}, \qquad k = 2,\ldots,n.$$

For a sample of size n from the interval [0, 1], the ratio ranges from 1/2 to (n-1)/n , cf. Table 9. This is rather wide variation

TABLE 9

The Ratio of Adjacent Order Statistics from the Modified Coleman Model

	Sample Size						
i	3	4	5	.	.	.	12
2	.667	.750	.800				.916
3	.500	.667	.750				.909
4		.500	.667				.900
5			.500				.889
6							.875
7							.857
8							.833
9							.800
10							.750
11							.667
12							.500

for a value which would be constant if the exponential relation held.

The conclusion is clear: if participation is exponentially distributed, it is not because the researchers have essentially ordered uniformly distributed random numbers and aggregated them.

Leik's Model

In 1965 Leik suggested that each person in a discussion group can be characterized by a tendency to speak, c_i, and that p_{ik}, the proportion of acts initiated by person i in group k, is a function of the tendencies present in the group; i.e.

$$p_{ik} = \frac{c_{ik}}{\sum_j \delta_{jk} c_{jk}}$$

where $\delta_{jk} = 1$ if person j is in group k, and 0 otherwise (1965:58-59). Leik asserted that if his theory were correct, then for a single group, the exponential distribution of participation would exist if and only if the tendencies to participate were exponentially related (1965:59). Subsequently, Leik has suggested that the distribution of c_i appears to be "skewed normal" and that random sampling from this distribution and the operation of his theory could produce the exponential distribution of participation (1967:285).[13]

Unfortunately, Leik's model says nothing about the frequencies which are produced by the actors and it is the aggregated frequencies, not the aggregated proportions, which appear to be approximated by the exponential function. Thus, for his model to be made relevant, some set of assumptions about the relation of the c_i and the frequencies is necessary.

Assume that c_i, the "tendency to participate," is a rate of participation, and let $c_i = f_i/t$, where f_i is the frequency by person i and t is the duration of the group. If groups within a particular study are of equal duration, then picking a sample from c_i is equivalent to picking a sample from f_i.

Rather than evaluate the extended Leik model directly, we shall evaluate a general model of which it can be shown the extended Leik model is a special case.

The General Model

Both the modified Coleman explanation and the extended Leik explanation rest on the aggregation of ordered observations chosen randomly from some type of distribution: uniform for Coleman and unimodal asymmetric for Leik. Although we have already shown that the modified Coleman model is not correct, the similarity of the two explanations suggests the more general question of whether there exists any distribution such that when random samples of size k are drawn, ordered, and aggregated, an exponential series will result.

Kadane (1968) has shown that if there is any distribution such that the order statistics from random samples of size k result in an exponential relation, then the order statistics from random samples of size j, $j \neq k$, will not produce an exponential relation (except in the degenerate case where the ratio is equal to 1). For random samples of size $j < k$ the ratio of the ith smallest to the ith + 1 smallest is monotonically increasing. Thus Leik's model is impossible; there is no distribution such that the required method will produce exponential series for more than one sample size.

This result, however, does not tell us whether there is a distribution such that for some sample size the required exponential relation would obtain. Further work has shown that there exists a set of distributions such that random sampling would result in the exponential relation for samples of size 3, there exists a set of distributions which do the same for samples of size 4, and there exists a set for samples of size 5 (cf. Kadane, 1968). It may be the case that in general there is, for every sample size, some set of distributions such that the specified procedure produces the exponential relation. It should be made clear, however, that the results obtained so far prove only that in these cases the required set of distributions is not empty; they do not give a convenient characterization of these distributions.

Nevertheless, a model which requires that in order to produce an exponential relation for different sample sizes there must be different sets of distributions, hardly constitutes an adequate explanation of the alleged regularity. This explanation would be relevant only if groups of different sizes had been picked from different populations, but in at least one instance (Stephan and Mishler, 1952) the researchers selected groups of different sizes from a single population.

An Approximation to the Exponential Relation

To this point, this alleged explanations and the discussion of them have been predicated on the assumption that the observed distributions of participation are characterized by exponential relations. It is quite obvious from the data that have been offered, however, that this characterization is only approximately true. One might reasonably ask whether there are distributions such that the specified method (random sampling, ordering, and aggregating) will produce results which are approximately exponential.

As the sample size gets large, the ratio of expected values of adjacent order statistics approaches 1 for any distribution. The difference between that ratio and 1 , divided by the sample size, approaches a number which may depend on which part of the distribution is being examined. The Appendix shows that a density proportional to $1/x$ has the property that the above number does not depend on which part of the distribution is being examined. Thus, a density proportional to $1/x$, for $0 < a \leq x \leq b < \infty$, would be expected, for large samples, to have nearly constant ratios of expected values of order statistics. To see how close to constant the ratios are for small samples, expected values of order statistics were computed for the density $1/(x \cdot \ln(15))$ for $1 \leq x \leq 15$. The ratios of adjacent expectations of order statistics, ranked highest to lowest, and the expectations as percentages are shown in Tables 10 and 11.

In Table 10 the ratios of expectations are shown for samples of size 3 through 12. As expected, the ratios are more nearly constant for larger sample sizes and the average ratio increases with sample size. For a given sample size, the ratios are monotonic decreasing. This result is similar to a result of the modified Coleman model (cf. Table 9), although the range of the ratios in the present case is only a fraction of the range in the modified Coleman model. Indeed, although the ratios are not constant, they appear to be at least as constant as the ratios for the observed data (cf. Table 7).

In Table 11 the expectations of order statistics are reported as percentage distributions. A significant fact in Table 11 is that the percentage accounted for by the largest expectation is a monotonic decreasing function of the sample size.

Any attempt to fit the exponential to the entire group and to groups of various sizes must take account of the fact that if the distributions have the same sum (e.g., 1.1 or 100), then it cannot simultaneously hold that r is a monotonic increasing function of n and p_1 is a constant. Only one of these facts could be true of groups in which the distribution of participation was characterized by an exponential relation. The implications of this are important in the present case as will be shown.

Assuming that one could give a meaningful interpretation to sampling from the density $1/x$, the result indicate that the required method produces ratios which are as constant as those for any of the data which have been reported, but r is a monotonic increasing function of n.

TABLE 10

The Ratio of Adjacent Order Statistics from the Approximation Model

	Sample Size				
i	3	4	5	6	7
2	.540	.614	.666	.704	.735
3	.497	.590	.651	.695	.728
4		.560	.634	.683	.720
5			.612	.670	.711
6				.655	.702
7					.690
$\frac{n}{n+\ln(15)}$*	.526	.596	.649	.689	.721

	Sample Size				
i	8	9	10	11	12
2	.759	.780	.797	.811	.824
3	.754	.776	.794	.809	.822
4	.749	.772	.790	.806	.820
5	.743	.767	.787	.804	.818
6	.736	.762	.783	.801	.815
7	.728	.757	.779	.798	.813
8	.720	.751	.775	.794	.810
9		.744	.770	.791	.808
10			.765	.787	.805
11				.783	.802
12					.799
$\frac{n}{n+\ln(15)}$	.747	.769	.787	.802	.816

*Cf. Appendix.

TABLE 11

Percentage Expectations from the Approximation Model

	Sample Size				
i	3	4	5	6	7
1	55.29	45.90	39.35	34.49	30.73
2	29.86	28.17	26.19	24.30	22.58
3	14.85	16.62	17.05	16.88	16.44
4		9.31	10.80	11.54	11.84
5			6.61	7.73	8.42
6				5.07	5.91
7					4.08

	Sample Size				
i	8	9	10	11	12
1	27.72	25.26	23.20	21.46	19.97
2	21.05	19.69	18.49	17.41	16.45
3	15.88	15.28	14.67	14.08	13.52
4	11.89	11.79	11.60	11.36	11.08
5	8.83	9.04	9.13	9.12	9.06
6	6.50	6.89	7.15	7.31	7.39
7	4.73	5.22	5.57	5.83	6.01
8	3.40	3.92	4.32	4.63	4.87
9		2.92	3.33	3.66	3.93
10			2.55	2.88	3.16
11				2.26	2.54
12					2.03

In Stephan and Mishler's data (including the formal leaders) and Bales et al.'s data, the persons who rank highest in units initiated produced about 44 percent, and there was no noticeable trend across various size groups. Thus, if the distribution of participation for all group members is to be approximated by the exponential relation, there would be a monotonic <u>decreasing</u> relation between the average ratio and the size of the group. Since this is contrary to the relation actually produced by sampling from the density $1/x$, the approximation model also must be rejected.

<u>Discussion</u>

None of the alleged explanations has turned out to be satisfactory; either (1) they did not imply that the distribution of participation would be exponential (Coleman and Leik), (2) they involved assumptions contrary to fact (the general sampling model), or (3) they implied additional consequences which were unsupported (the argument based on $1/x$.) Thus, stripped of these potential explanations, the only basis for adopting the exponential over the harmonic functions is the fit of the function to the data, and this appears to vary with type and size of the group.

In the face of the previous results, it seems reasonable to suggest that one forgo further explanations of the fit of the exponential and further curve fitting and, rather, consider the broader problem of developing models which could capture the process of interaction, and not just its end result. Whatever regularities occur as an outcome of the entire operation of the interaction process must also be capable of being generated from assumptions about the nature of the process itself. Some work and some analyses have begun along this line (cf. Bales, 1953; Horvath, 1965; Kadane et al., 1968; and Lewis, 1968).

In general, models allow one to derive and to specify the relations among multiple characteristics of the phenomenon modeled. Horvath's model of participation in discussion groups, for example, allows one to derive not only the exponential distribution of participation among actors, but also (1) the probability of k consecutive acts by a person of rank j, (2) the mean number of consecutive acts by a person of rank j, (3) the mean number of acts before the person of rank j speaks, given that the person of rank i is currently speaking, and so on (cf. Lewis, 1967).

The utility of being able to specify relations among various descriptions is well displayed in the present instance. Although the exponential function seems to fit Stephan and Mishler's data a lettle better than Bales', the claim had been made that the exponential function fit both data sets. Since Stephan and Mishler's unit of observation is very close to an uninterrupted sequence of Bales' units (cf. Kadane, et al., 1968), it is interesting to consider whether the same type of function <u>should</u> be expected to fit for both types of data. Alternatively, under what circumstances should both sets of data be characterized by an exponential function? As suggested by the preceeding comments, the answer to this question requires a model. In the case of Horvath's model, the answer is that the two distributions are never both characterized by exponential relations.

Finally, the type of data analysis employed by Bales et al. and by Stephan and Mishler has been employed by others. Coleman (1964:30) cites research on sociometric choices, sales of popular records, and sales of best selling books, as examples of research in which data have been ordered and aggregated and in which exponential relations allegedly have been found. Hopefully, the form of analysis and the analytic results in the present paper will be useful in evaluating those claims, and claims of a similar type in the future.

Summary

Sixteen years ago the claim was made that the exponential function described the distribution of ordered, aggregated data on participation in discussion groups. Reconsideration suggests that the claim is not as unequivocal as it has apparently been believed: (1) the harmonic function does nearly as well as the exponential function for some data, and (2) several potential explanations for exponential (or nearly exponential) relations were shown to be insufficient. Because many functions would probably fit as well as the functions which have been proposed, thought should be given to the construction of process models rather than to futher attempts at curve fitting.

Appendix: Distribution Functions with Order Expectations Which Approach an Exponential Series*

Let $o < X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics from an arbitrary continuous distribution function F, with $F(0) = 0$. Then we wish to find conditions on F so that $E[X_{(i+1)}]/E[X_{(i)}]$ does not depend on i.

$$\frac{E[X_{(i+1)}]}{E[X_{(i)}]} = 1 + \frac{\dfrac{n^{-1}(E[X_{(i+1)}] - E[X_{(i)}])}{n^{-1}}}{E[X_{(i)}]}$$

$$\sim 1 + \frac{n^{-1}\dfrac{dX_\alpha}{d\alpha}}{X_\alpha}$$

$$= 1 + n^{-1} \cdot \frac{d\ln X_\alpha}{d\alpha}$$

But this is the same as saying that

$$1 + n^{-1} \cdot \frac{1}{F^{-1}(\alpha) \cdot f(F^{-1}(\alpha))}$$

is not to depend on α. Or, letting $y = F^{-1}(\alpha)$, $y \cdot f(y)$ is not to depend on y. So that, locally,

$$f(y) \propto 1/y .$$

Thus the Pareto distribution with parameter 1 (truncated so that it is the density of a probability distribution) is an approximation to a density having the required property.

* The authors wish to thank I. R. Savage for helpful conversations which led to this argument.

FOOTNOTES

1. The use of chi square in this instance requires the assumption that the acts are independent.

2. The use of maximum likelihood estimates requires a stochastic model which determines the distribution of the observations. Some models of this type are discussed in Kadane et al. (1968).

3. Coleman recognized that a is determined in the special case where $C = 1.0$ (1960:53; 1964:29).

4. In the original report $\hat{r}$ for groups of size 6 should have been .603 rather than .623; and for groups of size 8 the percentage distribution on which the computation of $\hat{r}$ was based summed only to 99.2% . When $\hat{r}$ is computed from the recalculated percentage distributions, the values of $\hat{r}$ are not a strictly increasing function of n . This casts some doubt whether r is a linear function of n (cf. Stephan and Mishler, 1952:604-605, and Coleman, 1960:65-69).

5. Bales et al. (1951) included only the graphs and only for groups size three through eight. The frequency distributions on which the information in Table 2 is based were supplied to Coleman by Bales. The authors wish to express their appreciation to Robert Bales and James Coleman for making these data available.

6. It is important to point out that whereas Stephan and Mishler omitted the data from the instructor apparently for a priori reasons, we have omitted data from the highest participant solely to see how much it improves the fit to the exponential function.

7. The authors wish to thank Philip Bonacich for making these data available.

8. It is not necessary, however, that one aggregate exponential and non-exponential series for the resulting series to be non-exponential. One can show that if any two exponential series with the same number of elements are aggregated, the resulting series is exponential if and only if the two original distributions have the same parameter r. This result raises questions about the condidtions under which one wants to aggregate data.

9. The authors wish to thank Fred L. Strodtbeck and Margaret Parkman Ray for providing these data from unpublished work done under Family Interaction Studies Using Revealed Difference, NIMH Grant #MH05572-03 , in the Social Psychology Laboratory, at The University of Chicago.

It is interesting that if one aggregates the data according to the role of the participants one finds for the fathers, mothers, and children 34.8% , 33.0% , 32.2% . This is very close to the distribution reported by Leik (1965:64).

10. The authors wish to thank Mrs. Michael Olmsted and Ted Mills for making the four person Harvard data available and Zvi Namenwirth and Michael Farrell for making the four person Yale data available. For a description of the Harvard groups cf. Olmsted (1954).

11. Although it was not a necessary feature of his model, Coleman did in fact simulate all groups for the same number of acts.

12. If one expresses the expected distribution for samples of size five as percentages, this distribution is very close to the observed percentage distribution of participation for groups of five <u>females</u>, cf. Table 4, the mean error between these two distributions being .26% . Leik reports that the data from his groups which were composed of <u>females</u> "are frequently contradictory to the Stephan-Mishler hypothesis" (1965:59).

13. By "skewed normal" Leik intended "that the distribution be what would be obtained if, for example, the abscissa of a normal distribution were stretched or contracted by, e.g., a log transformation" (Letter from Leik, Feb. 1968). This appears to include not just log normal distributions but all asymmetric unimodal distributions.

REFERENCES

Bales, R.

1953 "The equilibrium problem in small groups." Pp. 111-161 in T. Parsons (ed.) et al., Working Papers in the Theory of Action. Glencoe, Ill.: The Free Press of Glencoe, Illinois.

_________ et al.

1951 "The channels of communication in small groups." American Sociological Review 16:461-468.

Bonacich, P.

1968 Specialization and Differentiation in Small Laboratory Groups. Ph.D. Dissertation. Harvard University.

Coleman, J.

1960 "The mathematical study of small groups." Pp. 1-150 in H. Solomon (ed.), Mathematical Thinking in the Measurement of Behavior. Glencoe, Ill.: The Free Press of Glencoe, Illinois.

____

1964 Introduction to Mathematical Sociology. N.Y.: The Free Press of Glencoe.

Horvath, W.

1965 "A mathematical model of participation in small group disucssions." Behavioral Science 10:164-166.

Kadane, J.

1968 "A moment problem for order statistics: General theory and an application." In preparation.

__________ et al.

1968 "Stochastic theories of participation in group discussions." Submitted for publication.

Leik, R.

1965 "Type of group and the probability of initiating acts." Sociometry 28:57-65.

____

1967 "The distribution of acts in small groups." Sociometry 30:280-299.

Lewis, G.

1967 "The consequences of formal theory." Paper read at the annual meeting of the American Sociological Association, August 28-31, San Francisco, Calif.

____

1968 "An analysis of Bales' Monte-Carlo model of interaction." In preparation.

Olmsted, M.

1954 "Orientation and role in the small group." American Sociological Review 19:741-5751.

Stephan, F.

1952 "The relative rate of communication between members of small groups." American Sociological Review 17:482-486.

__________ and E. Mishler

1952 "The distribution of participation in small groups: An exponential approximation." American Sociological Review 17:598-608.