SELECTING THE MOST RELIABLE FROM A SET OF JUDGES: TOWARD A GENERAL SOLUTION

Sandra L. Aivano, Yale University
Domenic V. Cicchetti and Jacob Levine, V.A. Hospital, West Haven, Conn.

Introduction

Social scientists frequently study variables that can be measured only by means of ordinal rating scales. Since the quality of the data collected using such scales could greatly influence the results obtained, pilot tests are often run to insure that the ratings in the main investigation will be accurate and reliable. Within this context, researchers often face the task of selecting the most reliable judges or raters to participate in the study.

One pilot testing procedure frequently used to assess raters' reliability and to choose the most reliable among them is to compare the raters' judgments to those of an expert or experts (Lehmann, Ban, and Donald, 1965; Fleiss, Spitzer, and Burdock, 1965). Judges whose scores deviate greatly from the expert judgments are either eliminated or trained in the use of the rating scale. This technique is not always feasible, since it assumes that expert judges can be found for the variable in question and that they can agree among themselves. However, it is difficult to establish experts on the basis of training or experience for many variables (such as the one to be discussed in this paper). Furthermore, it is often impossible or impractical for experts to participate in reliability studies (Fleiss, Spitzer, and Burdock, 1965; Smith, 1974). In view of the problems with this procedure, it is clear that a method for choosing reliable judges is needed that does not depend on expert judgments.

Several techniques have been proposed for comparing the reliability of judges and identifying the most reliable among them. One method involves computing the intraclass correlation coefficient (R_I) for all subsets of R raters from the pool of R′ raters, and selecting the subset with the highest R_I value (Burdock, Fleiss, and Hardesty, 1963). Another technique is to compute the Spearman Rho between each rater's ranking of the subjects and a composite ranking reflecting the rankings by all the remaining raters (Smith, 1974).

A new technique for selecting the most reliable raters from a larger rater pool is presented in this paper. This method was developed in the context of a pilot study designed to establish the difficulty of a series of cartoons to be used in a later investigation. Preliminary analyses revealed that agreement on this variable was not, in general, above chance expectancy. Thus, we were faced with the problem of determining which raters from the initial group could rate the cartoons reliably. This technique bears similarities to a technique briefly mentioned by Smith (1974) of computing the intercorrelations between all possible rater pairs, converting these values using Fisher's Z function, and using the average Z value for each rater as an index of that rater's reliability relative to the other prospective raters. Rather than using intercorrelations, however, the method presented here employs weighted kappa due to Cohen (1968) with a standard error due to Fleiss, Cohen, and Everitt (1969). Weighted kappa is specifically designed to measure agreement when the data are ordinal. It is more appropriate than other measures of association for ordinal scales since it takes into consideration the amount of agreement expected by chance alone. (For a further discussion of the kappa statistics relative to other available statistics for assessing rater agreement with qualitative data, see Fleiss (1975).)

Method

Ten judges rated the difficulty of 30 cartoons selected from magazines. The raters were staff psychologists at the V.A. Hospital, West Haven, Connecticut and Yale University. The variable, Difficulty Level, was measured on a 5-point ordinal scale with the following categories: (1) "very easy"; (2) "easy"; (3) "average"; (4) "difficult"; and (5) "very difficult."

Results and Discussion

Agreement Statistics

The agreement for each of the R(R-1)/2 rater pairs was assessed using weighted kappa (Cohen, 1968; Fleiss, Cohen, and Everitt, 1969) with a continuous-ordinal weighting system (Cicchetti, 1972, 1976; Cicchetti and Allison, 1973) as defined below. Weights (W) were computed by the formula:

$$W = \frac{k-1}{k-1}, \frac{k-2}{k-1}, \ldots \frac{k-(k-1)}{k-1}, \frac{k-k}{k-1} \qquad [1]$$

where k refers to the number of points on the scale. Weights range from $\frac{k-1}{k-1}$ or 1, for ratings in perfect agreement, to $\frac{k-k}{k-1}$ or 0, for those that are the maximum possible number of scale points apart. Weights for the various levels of partial agreement, $\frac{k-2}{k-1}, \ldots \frac{k-(k-1)}{k-1}$, assume values between 0 and 1. Using this weighting system, the proportion of observed agreement (PO); proportion of expected or chance agreement (PC); the level of chance-corrected agreement, or kappa, i.e., (PO-PC)÷(1-PC); the Z value of kappa; and its level of statistical significance, were computed for each rater pairing.

A recent Monte Carlo study has revealed that a minimum sample size of approximately $2k^2$ is needed to obtain valid results with the kappa statistics (Cicchetti and Fleiss, 1976). Since the thirty cartoons do not constitute a sufficient sample for the 5-point scale of Difficulty Level, categories 1 and 2 and categories 4 and 5 were

collapsed, and the data were reanalyzed on a 3-point scale. The results obtained with the 3-point scale were very similar to those obtained using the 5-point scale. Ten and eleven statistically significant kappa values ($p \leq .05$) were obtained using the 3-point and 5-point rating systems, respectively. Nine of these significant kappas were for the same rater pairs on both the 3-point and 5-point scales. For the sake of brevity, the kappa statistics for only the 3-point scale are reported in Table 1.

Ranking Systems

As indicated from the kappa statistics in Table 1, the agreement for most judge pairs (35 from a total of 45) was not above chance expectancy at $p \leq .05$. Despite this overall lack of agreement, we wished to identify the raters that were the most reliable. To accomplish this, we developed two systems for ranking the raters based on the significance of the kappa values obtained in the pairwise comparisons.

Ranking System 1 consists of assigning consecutive integer ranks to the R(R-1)/2 rater pairings according to the magnitude of their Z of kappa values, with a rank of 1 for the rater pair with the highest Z value, and a rank of R(R-1)/2 for the rater pairing with the lowest Z value. The ranks for the R-1 comparisons associated with each rater are summed to obtain a composite rank reflecting that rater's reliability relative to the remaining raters. Then the composite rank scores of the raters can be compared, and the raters with the lowest scores can be identified as the most reliable. The results obtained by applying Ranking System 1 to the kappa statistics in Table 1 are presented in Table 2.

While System 1 is an index of the *relative* standing of each rater, it does not take into account the absolute level of the Z values. However, the magnitude of the Z values is often of great importance to the researcher, since it is of little value to identify the most reliable judges if, say, none of the kappa values obtained is statistically significant or all of the kappas are highly significant. Thus, to provide additional information for selecting the most reliable raters, we developed a second ranking system.

Ranking System 2 utilizes the number of significant ($p \leq .05$) and approaching significant ($p \leq .10$) Z of kappa values among the R-1 kappas associated with each rater. The raters are first ranked according to the number of significant Z values among the R-1 comparisons associated with each of them. Raters with the same number of significant Z values are further differentiated by the number of Z values approaching significance. The results obtained by applying this ranking system to the kappa statistics for the 45 rater pairings from Table 1 are presented in Table 3.

Selecting the Most Reliable Raters

As can be seen from Tables 2 and 3, the rank orderings of raters produced by the two ranking systems are very similar. The rank correlation coefficient (Spearman's Rho) between the two orderings is 0.83 with $p = .002$. The researcher is at liberty to decide how to divide the raters into subsets of the most reliable and least reliable, using the information obtained from the two ranking systems. For the 10 raters of Difficulty Level, we felt that Raters 2, 5, 6, 7, 8, and 9 could be considered reliable, while Raters 1, 3, 4, and 10 were markedly less reliable. This division of the raters seems reasonable, since it splits the raters at one of the points of greatest difference in composite scores (in Table 2, between 9 and 3 who are 34 points apart), and eliminates those raters with only one significant Z value (in Table 3). The *same* six raters (2, 5, 6, 7, 8, and 9) are identified as the most reliable by both ranking systems.

Both ranking systems proposed above evaluate the raters on the basis of their agreement *with all other raters*. Another important consideration in selecting the most reliable subset of judges is the extent to which the judges selected agree *among themselves*. Table 4 presents the proportion of observed agreement (PO); proportion of chance or expected agreement (PC); and the significance level (p) of chance-corrected agreement (or Kappa), for rater pairs in the most reliable subset (Raters 2, 5, 6, 7, 8, and 9) and least reliable subset (Raters 1, 3, 4, and 10), respectively. A comparison of the two portions (A and B) of the table shows that the overall levels of agreement between the more reliable raters are much higher than those between the less reliable raters. The mean of the PO values for the most reliable raters is 0.70 compared to a mean of 0.59 for the less reliable raters. Of the 10 significant ($p \leq .05$) kappas among all 45 rater pairs, 7 are between raters in the most reliable subset, while only 1 significant kappa is found in the least reliable subset. Thus, the raters who are most reliable with respect to all raters in the pool are also highly reliable with respect to each other.

The method employed here to find the most reliable subset of the ten raters of cartoon Difficulty is suggested as a general technique for selecting the most reliable judges from a larger judge pool. Comparing each rater relative to all the others using ranking systems based on the levels of chance-corrected agreement, as measured by Z of kappa values, seems a reasonable approach to the problem. This technique is based on the assumption that raters with the highest agreement relative to all the raters are, indeed, the best suited to participate in later investigations.

Another approach to selecting raters is to choose those with the greatest agreement among themselves, disregarding their levels of agreement with the remaining raters in the pool. This might be accomplished by computing the kappa statistics for all possible rater pairs and averaging the Z of kappa values for each rater pairing among the raters in each possible R rater subset of the R′ rater pool. This technique is often impractical, however, since the researcher must either decide

beforehand the size of the subset he wishes to select or he must compute average Z values for all the possible subsets of all sizes. Further, this technique could result in the selection of raters who might not be well suited for future studies. For example, raters who tended to give the highest possible score would agree highly among themselves, yet would probably not be using the scale properly.

A computer program has been written to implement the rater selection method discussed in this paper. For each possible rater pair the program computes *weighted* kappa statistics, if the data are *ordinal*, or *unweighted* kappa statistics (Cohen, 1960; Fleiss, Cohen, and Everitt, 1969), if the data are *nominal*. Further, if the sample size is sufficiently large, bias is assessed for each rater pair by means of the chi-squared statistic developed by McNemar (1947). Finally, rankings of the raters by the two ranking systems, as well as a number of summary tables, are also provided.

Summary

Researchers often face the task of identifying the most reliable subset of a given set of judges. Specifically, this problem arose when ten judges rated the difficulty of a series of cartoons on an ordinal scale. Kappa statistics and various ranking techniques were used to obtain the information necessary to select the most reliable judges from the original pool. The observed agreement, chance agreement, and the statistical significance of their difference were computed for each pair of judges. These data were summarized in several tables, using ranking systems for both the judge pairs and the individual judges. An important contribution of one interjudge ranking system is that it assigned to each judge a composite score reflecting the reliability of his ratings relative to that of all the others. Finally, a computer program was written for use in resolving this type of research problem.

Acknowledgments

This research was supported by the West Haven V.A. Hospital (MRIS 1416). The authors express their appreciation to Lorraine R. Gambino for her excellent typing of the manuscript and editorial assistance.

References

Burdock, E.I., Fleiss, J.L., and Hardesty, A.S. A new view of interobserver agreement. *Personnel Psychology*, 1963, *16*, 373-384.

Cicchetti, D.V. A measure of agreement between rank ordered variables. *Proceedings of the American Psychological Association*, 1972, *7*, 17-18.

Cicchetti, D.V. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 1976, in press.

Cicchetti, D.V. and Allison, T. Assessing the reliability of scoring EEG sleep records: An improved method. *Proceedings and Journal of the Electrophysiological Technologists' Association*, 1973, *20*, 92-102.

Cicchetti, D.V. and Fleiss, J.L. Comparison of the null distributions of Weighted Kappa and the C Ordinal Statistic. *Applied Psychological Measurement*, 1976, in press.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Fleiss, J.L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, *31*, 651-659.

Fleiss, J.L., Cohen, J., and Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Fleiss, J.L., Spitzer, R.L., and Burdock, E.I. Estimating accuracy of judgment using recorded interviews. *Archives of General Psychiatry*, 1965, *12*, 562-567.

Lehmann, H.E., Ban, T.A., and Donald, M. Rating the rater. *Archives of General Psychiatry*, 1965, *13*, 67-75.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947, *12*, 153-157.

Smith, J.M. A new rater selection technique for use with behavioral rating scales. *Journal of Clinical Psychology*, 1974, *30*, 40-43.

TABLE 1

AGREEMENT BETWEEN RATER PAIRS FOR 10 RATERS ON DIFFICULTY LEVEL FOR 30 CARTOONS ON A 3-POINT SCALE

Rank	Rater Pair[1]	PO	PC	Kappa	Z of Kappa	p of Kappa
1	3, 10	.77	.59	.43	2.94	.003
2	6, 8	.72	.61	.28	2.83	.005
3	2, 8	.78	.64	.40	2.80	.005
4	4, 7	.75	.60	.38	2.73	.006
5	5, 6	.75	.64	.31	2.68	.007
6	2, 9	.77	.63	.38	2.57	.010
7	2, 5	.72	.60	.30	2.12	.034
8	7, 9	.63	.52	.23	2.08	.037
9	7, 8	.65	.55	.22	2.05	.040
10	1, 9	.77	.67	.30	1.98	.048
11	8, 9	.75	.66	.27	1.86	.062
12	3, 8	.72	.62	.25	1.79	.073
13	5, 8	.70	.61	.22	1.65	.098
14	1, 6	.63	.57	.14	1.65	.099
15	2, 6	.67	.60	.17	1.63	.104
16	1, 7	.60	.52	.16	1.61	.108
17	3, 5	.68	.59	.23	1.60	.109
18	5, 9	.68	.60	.21	1.60	.109
19	2, 7	.63	.55	.19	1.59	.112
20	6, 9	.63	.57	.14	1.52	.128
21	2, 4	.58	.51	.15	1.38	.166
22	4, 8	.57	.50	.13	1.34	.179
23	5, 7	.65	.58	.17	1.33	.183
24	2, 3	.67	.60	.17	1.14	.253
25	6, 7	.70	.65	.13	1.09	.276
26	5, 10	.65	.59	.15	1.03	.302
27	1, 8	.72	.67	.14	.94	.346
28	1, 5	.65	.60	.12	.91	.361
29	1, 3	.67	.62	.12	.88	.376
30	3, 7	.60	.55	.11	.88	.376
31	4, 5	.60	.55	.10	.88	.378
32	3, 6	.63	.60	.08	.77	.438
33	6, 10	.63	.60	.08	.77	.438
34	4, 9	.52	.48	.07	.74	.457
35	8, 10	.65	.62	.08	.56	.578
36	4, 10	.55	.52	.06	.55	.581
37	3, 9	.63	.61	.06	.42	.676
38	7, 10	.57	.55	.04	.29	.768
39	4, 6	.65	.64	.02	.13	.895
40	2, 10	.60	.60	.00	.00	1.00
41	3, 4	.52	.52	-.01	-.06	.951
42	1, 2	.63	.64	-.02	-.11	.914
43	1, 10	.60	.62	-.05	-.38	.704
44	1, 4	.45	.47	-.04	-.46	.644
45	9, 10	.57	.61	-.11	-.78	.437

[1]Rater pairs are ordered by Z of kappa values.

TABLE 2

RANKING SYSTEM 1

RANKING THE RATERS BY COMPOSITE RANK SCORE

Rater	Composite Score	Ranks for Rater Pairs Comprising Composite Score
8	134	2+3+9+11+12+13+22+27+35
5	168	5+7+13+17+18+23+26+28+31
7	172	4+8+9+16+19+23+25+30+38
2	177	3+6+7+15+19+21+24+40+42
6	185	2+5+14+15+20+25+32+33+39
9	189	6+8+10+11+18+20+34+37+45
3	223	1+12+17+24+29+30+32+37+41
1	253	10+14+16+27+28+29+42+43+44
4	272	4+21+22+31+34+36+39+41+44
10	297	1+26+33+35+36+38+40+43+45

TABLE 3

RANKING SYSTEM 2

RANKING THE RATERS BY THE NUMBER OF SIGNIFICANT ($p \leq .05$) AND APPROACHING SIGNIFICANT ($p \leq .10$) Z OF KAPPA VALUES

Rater	$p \leq .05$	$.05 < p \leq .10$	$p \geq .10$
8	3	3	3
9	3	1	5
2	3	0	6
7	3	0	6
5	2	1	6
6	2	1	6
1	1	1	7
3	1	1	7
4	1	0	8
10	1	0	8

TABLE 4

WEIGHTED KAPPA STATISTICS FOR RATER PAIRINGS OF THE MOST RELIABLE AND LEAST RELIABLE RATERS: PO, PC, AND THE LEVEL OF STATISTICAL SIGNIFICANCE OF KAPPA (p)

A. Pairings of the Most Reliable Raters

Rater		2	5	6	7	8
5	PO	.72				
	PC	.60				
	p	.034*				
6	PO	.67	.75			
	PC	.60	.64			
	p	.104	.007**			
7	PO	.63	.65	.70		
	PC	.55	.58	.65		
	p	.112	.183	.276		
8	PO	.78	.70	.72	.65	
	PC	.64	.61	.61	.55	
	p	.005**	.098+	.005**	.040*	
9	PO	.77	.68	.63	.63	.75
	PC	.63	.60	.57	.52	.66
	p	.010*	.109	.128	.037*	.062+

B. Pairings of the Least Reliable Raters

Rater		1	3	4
3	PO	.67		
	PC	.62		
	p	.376		
4	PO	.45	.52	
	PC	.47	.52	
	p	.644	.951	
10	PO	.60	.77	.55
	PC	.62	.59	.52
	p	.704	.003**	.581

\+ = Significant at .10 level
* = Significant at .05 level
** = Significant at .01 level