1

DOCUMENT RESUME

ED 040 964 SP 004 013

AUTHOR	Rummery, Robert E.
TITLE	Characteristic Components of Interrater Variation in Judgments of Teaching Performance.
PUB DATE	70
NOTE	19p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.05
DESCRIPTORS	*Classroom Observation Techniques, *Reliability, *Statistical Analysis
IDENTIFIERS	Classroom Observation Judgment Schedule

ABSTRACT

A study concerned with identifying sources of interrater variation in ratings posed the following questions: Are ratings decomposable into a single nonerror component with interrater variations representing individual error components, or is a better fit to the data provided by multiple nonerror components representing generalized rating styles? And if multiple rating styles are found, what are their characteristics? Rated events were 10-minute segments from videotapes of high school classes in four different subjects. The 50-minute composite videotape was viewed by 83 subjects (teachers, teacher trainees, school administrators, and graduate students) using a 21-item questionnaire synthesized from a variety of sources to sample three aspects of teaching behavior: intended objectives, teaching style, and interpersonal climate. The data from ratings of the four classrooms with the 21 scales formed an 83 x 21 x 4 data array. Two analyses were performed on the extended matrix: principal component analysis of covariances and correlations between rows. Additional analytical procedures were employed to characterize generalized rating styles. Conclusions are methodological rather than substantive: The analytical procedures offer the possibility of providing more information about the quality of ratings than is provided by more traditional reliability estimation procedures, and provide a basis for selecting raters having rating styles of particular interest. (Observation schedule and data tables included.) (JS)

CHARACTERISTIC COMPONENTS OF INTERRATER VARIATION IN JUDGMENTS OF TEACHING PERFORMANCE

Robert E. Rummery

Illinois State University

INTRODUCTION

Despite critical commentary about the quality of information provided by ratings, they continue to be a popular source of data about classroom behavior--either as criteria of teacher effectiveness, or as indices of operative variables in the classroom situation. Ratings will undoubtedly continue to be widely used because they are easy and inexpensive to use and because they often provide abstractive information not readily available any other way. Claims that ratings are unreliable (Biddle, 1967) and that they may not measure what they are intended to measure (Guilford, 1962) suggest scrutiny of several aspects of rating methods, especially in instructional research. This paper deals with the specific question of identifying sources of interrater variation in ratings.

Before proceeding to a description of the problem and procedures for investigating it, a brief account of the genesis of the problem is in order. The starting point for the account is the assertion that ratings are unreliable. The statement is a troublesome one: the term "reliability" is used ambiguously; and the assertion is, in large part, undocumented. Strictly speaking, a necessary condition for estimating reliability of ratings is that a set of raters rate a common set of events. Estimation of reliability of ratings as stability would require that a set of raters make repeated observations of the same set of events; but such conditions are rarely available.

Two empirical approaches predominate in estimating reliability as equivalence. The first, restricted to multi-item rating devices, is internal consistency estimation; the preferred procedure probably being intraclass correlation or other analysis-of-variance-based procedure. The second, usually but not necessarily restricted to rating devices producing a single score, involves treating multiple raters as analogous to equivalent test forms. In both approaches, decomposition of ratings into independent components is on the basis of the classical test theory model

$$x_{ers} = t_{ers} + e.$$

LaForge (1965) has pointed out that in the multiple rater situation, there may be more than one way to relate ratings to patterns of behavioral cues. The classical model essentially takes into account only the most popular view; when, in fact, minority views may be just as relevant and just as free of error.

LaForge's article suggested as an alternative that individual ratings might be decomposable into r independent nonerror components, each one representing a different way of mapping patterns of cues into ratings--a different "rating style." The choice of a best-fitting decomposition model is empirically testable. If the classical model provides the best fit, the principal components of a matrix of rater intercorrelations will be found to consist of one component with a large characteristic root and k - 1 components with much smaller, approximately equal characteristic roots. If multiple rating styles are represented in the data, rater intercorrelations will produce two or more components with large characteristic roots. Determination of the meaning of "large" will be dealt with later.

LaForge's argument is consistent with Remmers' (1963) argument that ratings are the output of perceptual processes. Remmers' argument may be extended by considering ratings as responses functionally related to objective properties of observed events and to internal perceptual mechanisms of individual raters. Differences between raters in internal perceptual mechanisms could be represented as differences in parameter values of functional relationships between event-properties and perceptual output. This argument suggests the relevance of Tucker's work (1958, 1966) in the use of principal component analysis in the determination of parameters of functional relationships. Since one of the parameters might well be associated with individual differences in the dispersion of ratings, either over scales or over event, principal component analysis of covariance matrices also represents an appropriate basis for identification of generalized rating styles.

The present study can be considered as an extension of the LaForge study. The basic question is the same: are ratings decomposable into a single nonerror component with interrater variations representing individual error components or is a better fit to the data provided by multiple nonerror components representing generalized rating styles? An additional question is posed: if multiple rating styles are found in a set of rating data, what are the characteristics of the multiple rating styles? This study differs from the LaForge study in three other respects: the rated events were videotaped segments of secondary school classes, the ratings themselves were vectors of scores on multiple scales rather than single score ratings, and some additional analytical procedures were employed to characterize generalized rating styles.

PROCEDURE

The rated events in the study were ten-minute segments from videotapes of four classes recorded at University High School in Normal, Illinois. Ten-minute segments from classes in World History, Chemistry, General Mathematics, and American History were combined into a 50-minute composite allowing three-minute pauses between segments. The composite videotape was viewed by 83 subjects--24 teachers trainees, 22 classroom teachers, 21 school administrators, and 19 graduate students enrolled either in guidance or school psychology programs.

The rating device was a 21-item questionnaire synthesized from a variety of sources to sample three aspects of teaching behavior referred to by Sorenson and Gross (1965): intended objectives of instruction, teaching style, and interpersonal climate. Seven items were intended to convey information about elements of a subject-matter mastery orientation; seven were related to interpersonal climate; and seven were intended to characterize teaching styles between the extremes of didactic teaching and discovery teaching. A copy is included in the Appendix.

The data from ratings of the four classroom behavior samples with the 21 scales formed an 83 X 21 X 4 data array. Analysis proceeded on the extended two-way array of 83 row supervectors of four 21-element vectors (Horst, 1965. Pp. 317-324.). Two analyses were performed on this extended matrix: principal component analysis of covariances between rows, and principal component analysis of correlations between rows. Analysis of the covariance matrix permitted more detailed analysis of generalized rating styles. In addition, the analysis of the covariance matrix produced a reduced matrix of projections of scale-classroom combinations on the principal components. Unfolding analysis of order relations among these coefficients provided further information about characteristics of rating styles.

RESULTS

The characteristic roots of the covariance matrix are presented in Table 1 of the Appendix, along with increments between successive roots, variance accounted for by the component associated with each root, and the cumulative variance associated with successive components. The same information obtained from analysis of the correlation matrix is presented in Table 2 of the Appendix. At this point, the question of how many nonerror components best characterize the data arises.

LaForge cited two criteria for deciding how many components to retain. The first criterion, a psychometric one, indicates retaining all components associated with characteristic roots with values greater than one. For the correlation matrix, this criterion would result in the retention of 19 components. For the covariance matrix this criterion is meaningless since the disperions of individual ratings are not standardized. The second criterion involves a statistical test of differences in magnitudes of successive roots. The statistical criterion was not applicable for this particular correlation matrix because the value of the determinant, required in making the test, was approximately zero. The determinant of the covariance matrix was not obtained.

Another criterion has been suggested by Gulliksen (1959), related to the asymptotic nature of a plot of the magnitude of characteristic roots as a function of their ordinal number. Application of this criterion indicates retention of two components of the correlation matrix and three of the covariance matrix. The difference in the number of factors between the covariance matrix and correlation matrix reflects the fact that interrater variations in dispersion of ratings are retained in the covariance matrix, but not in the correlation matrix.

Loadings of individuals on the principal components of the correlation matrix are presented in Table 3 of the Appendix. These loadings represent correlations of ratings of individual raters with what may be interpreted as the true scores for generalized rating styles. The first three components of the covariance matrix accounted for approximately 45 percent of total variance; the first two components of the correlation matrix accounted for approximately 40 percent of total variance. The variance accounted for by the first component of the covariance matrix was approximately 29 percent as compared to about 30 percent for the correlation matrix; hence, a substantially better fit is provided by the representation of multiple nonerror components. The large amount of random variation remaining may be due to the fact that only four events were rated with the 21 scales, attenuating variance of individual scales over events.

The coefficients of the 84 classroom-scale observation units for the three principal components were represented in three 21 X 4 tables. The three 21 X 4 tables are combined in Table 4 of the Appendix. Each row of each of the three tables generates a rank ordering of the four classroom segments on a single scale. The orderings can be interpreted as representing an order of proximity to the ideal point of a scale for a rater utilizing each generalized rating style. This interpretation suggests the applicability of unfolding analysis (Coombs, 1964) for representation of the characteristics of the generalized rating styles. The existence of six rankings of a set of four objects (I-scales) unfoldable into a single rank order and its mirror image (a J-scale) provide the basis for inference of a single attribute underlying the six rankings. The existence of more than one set of six unfoldable orders allows the inference of additional attributes. The orders of the four classroom segments associated with the three components and the J-scales recovered from these orders are presented in Tables 5, 6, and 7 in the Appendix.

For the first component, rankings of the four classroom segments produced two J-scales. The first J-scale, defined by the order BDAC and its mirror image CADB suggests a contrast between careful preparation, clear organization, and intergration of topics to inattentiveness of students, deficiency in scholarship, and fault-finding and unfriendliness in the classroom. The second J-scale, defined by the order DCBA and its mirror image ABCD, is interpreted as a contrast between acceptance of pupil's ideas and permissiveness and teacher determination of topics and teacher involvement with the whole class in contrast to small groups of pupils.

For the second and third components, ranking of the classroom segments was predominantly unidemensional. For the second component, the ordering attribute is represented by a J-scale defined by the order CDAB and its mirror image BADC. For the third component, the ordering attribute is represented by a J-scale defined by the order CADB and its mirror image BDAC. Although noncollinear with the second J-scale recovered from the first component, the J-scale recovered from the second component was indistinguishable from it. The unfolding set recovered from the third component was incomplete but suggested a contrast between superior scholarship and teacher dominance of the classroom.

DISCUSSION

The conclusions to be reached from the investigation reported here are methodological rather than substantive. In tha data obtained, it is clear that individual ratings were decomposable into more than one non-error component, but no claim is made that these results would generalize to another sample of raters, another set of rating scales, or another set of events. The analytical procedures offer the possibility of providing more information about the quality of ratings than is provided by more traditional reliability estimation procedures and provide a basis for selecting raters having ratings styles of particular interest, as suggested by Anderson and Hunka (1964). The interpretations of the generalized rating styles are somewhat tentative because of the small number of events observed. Work is underway to compare the results of this form of analysis to the results of reliability estimation based on analysis of variance of the events by scales by raters classification. In addition, production of additional videotapes is underway to provide a larger number of events leading to a more adequate characterization of individual rating styles.

REFERENCES

Anderson, C. C. & Hunka, S. M. Teacher evaluation: Some prolbems and a proposal. Harvard Educational Review, 1963, 33, 74-96.

Biddle, B. J. Methods and concepts in classroom research. Review of Educational Research, 1967, 37, 337-357.

Guilford, J. P., Christensen, P. R., Taaffe, G., & Wilson, R. C. Ratings should be scrutinized. Educational and Psychological Measurement, 1962, 22, 439-447.

Gulliksen, H. Mathematical solutions for psychological problems. American Scientist, 1959, 47, 178-201.

Horst, P. Factor Analysis of Data Matrices. New York: Holt, Rinehart & Winston, 1965.

LaForge, R. Components of reliability. Psychometrika, 1965, 30, 187-195.

Remmers, H. H. Rating methods in research on teaching. In N. L. Gage (Ed.). Handbook of Research on Teaching. Chicago: Rand McNally, 1963. Pp. 329-378.

Sorenson, G. & Gross, C. F. Teacher appraisal: A matching process. Unpublished manuscript. University of California at Los Angeles, 1966. Mimeo.

Tucker, L. R. Determination of parameters of a functional relation by factor analysis. Psychometrika, 1958, 23, 19-23.

Tucker, L. R. Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.). Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally, 1966. Pp. 476-501.

APPENDIX

CLASSROOM OBSERVATION JUDGEMENT SCHEDULE

Observer______________________________ Class________________________

1. Teacher's preparation for class meeting.	no evidence of preparation	moderately well prepared	very carefully prepared
2. Teacher's ability to arouse pupil's interest.	majority of pupils inattentive	pupils mildy interested	pupil's interest very hi
3. Teacher's organization of instructional material.	no sign of system or order	some organization apparent	organization clearly apparent
4. Topic emphasis; balance between fundamentals and trivia.	neglect fundamentals for trivia	half fundamentals; half trivia	stresses fundamentals; disregards trivia
5. Scholarship; knowledge of subject matter.	clearly deficient	textbook competency	clearly superior
6. Ability to express ideas.	inarticulate; obscure	rather hesitant; slightly obscure	fluent; clear
7. Integration of lesson topics.	lesson topics isolated	some integration of lesson topics	all topics integrated
8. Acceptance of pupils' ideas	rejects all pupil ideas	accept ideas having merit	accepts all pupils' ideas
9. Acceptance of pupils' behavior.	highly critical	critical of extreme deviancy	highly permissive
10. Attitude toward pupils.	unsympathetic; inconsiderate	generally somewhat considerate	courteous and considerate
11. Social distance from pupils.	faultfinding; unfriendly	serious; somewhat reserved	conversational; friendly
12. Formality of classroom procedures.	rigidly formal structured	rather informal; somewhat structured	informal unstructured

PLEASE TURN PAGE

	Low end	Middle	High end
13. Manifest anxiety in classroom.	highly tense; anxious	generally relaxed; some tension	no sign of anxiety
14. Discipline and order in classroom.	order strictly maintained	some disorder but no nonsense	pupils self-regulating
15. Verbal output initiated by teacher.	10%	50%	90%
16. Relative information contribution of teacher.	10%	50%	90%
17. Size of classroom group(s) with which teacher is involved.	1 or 2 pupils	half of class	nearly all of class
18. Degree of teacher involvement with group(s).	minimal involvement	involvement limited to guidance	active participation in all groups
19. Determination of topics to be considered.	determined by class interests	teacher determination modified by class interests	total teacher determination
20. Task focus.	focus on critical analysis of sources of facts	some critical analysis of sources of factual content	focus on factual content
21. Inductive-deductive focus of class.	topic sequence from facts to generalization	facts and generalizations in no sequence	topic sequence from generalization to specific facts

TABLE 1

Characteristic Roots of Covariance Matrix

k	Root X 10^{-4}	Increment $(\lambda_{k+1} - \lambda_k)$	Percent of Variance	Cumulative Percent of Variance
1	9.891	--	29.91	29.91
2	3.555	6.336	10.75	40.66
3	1.512	2.043	4.57	45.23
4	1.367	.145	4.13	49.36
5	1.252	.115	3.79	53.15
6	1.135	.117	3.43	56.58
7	.924	.211	2.80	59.38
8	.801	.123	2.42	61.80
9	.783	.018	2.37	64.17
10	.709	.074	2.14	66.31
11	.674	.035	2.04	68.35
12	.599	.075	1.81	70.16
13	.582	.017	1.76	71.92
14	.554	.028	1.68	73.60
15	.496	.058	1.50	75.10
16	.452	.044	1.37	76.47
17	.440	.012	1.33	77.80
18	.437	.003	1.32	79.12
19	.392	.045	1.19	80.31
20	.384	.008	1.16	81.47
21	.357	.027	1.08	82.55

TABLE 2

k	Root	Increment $(\lambda k + 1 - \lambda k)$	Percent of Variance	Cumulative Percent of Variance
1	23.732		28.59	28.59
2	9.403	14.329	11.47	39.92
3	3.768	5.635	4.54	44.46
4	3.245	.523	3.91	48.37
5	3.017	.228	3.64	52.01
6	2.828	.189	3.40	55.41
7	2.297	.531	2.77	58.18
8	2.065	.232	2.49	60.67
9	1.988	.077	2.41	63.06
10	1.848	.140	2.23	65.29
11	1.572	.276	1.91	67.18
12	1.560	.012	1.88	69.06
13	1.458	.102	1.76	70.32
14	1.403	.055	1.69	72.51
15	1.229	.174	1.48	73.99
16	1.192	.037	1.44	75.43
17	1.128	.064	1.36	76.79
18	1.073	.055	1.29	78.08
19	1.034	.039	1.24	79.32
20	.979	.055	1.18	80.50
21	.945	.034	1.14	81.64

TABLE 3

Factor Loadings of Raters on Principal

Components of Correlation Matrix

Rater	I	II	Rater	I	II
1	.756	-.209	43	.656	-.215
2	.770	-.239	44	.646	-.277
3	.650	-.298	45	.670	-.235
4	.669	-.271	46	.681	-.099
5	.641	-.383	47	.397	-.382
6	.499	-.388	48	.764	-.249
7	.439	.023	49	.736	-.076
8	.310	-.390	50	.645	-.146
9	.704	-.208	51	.658	-.329
10	.670	-.211	52	.623	-.129
11	.535	-.368	53	.733	-.033
12	.770	-.251	54	.368	-.064
13	.644	-.181	55	.670	-.183
14	.739	-.211	56	.655	-.116
15	.452	-.077	57	.497	-.076
16	.526	-.038	58	.623	-.334
17	.222	.468	59	.617	-.296
18	.699	-.168	60	.515	-.216
19	.522	-.221	61	.718	-.274
20	.357	.248	62	.762	-.296
21	.378	.319	63	.767	-.266
22	.560	.447	64	.461	-.273
23	.588	.210	65	.325	.439
24	.482	.291	66	.596	.269
25	.304	.545	67	.459	.480
26	.357	.348	68	.210	.384
27	.268	.361	69	.066	.100
28	.404	.552	70	.262	.512
29	.443	.452	71	.210	.463
30	.344	.587	72	.718	-.282
31	.386	.361	73	.241	.555
32	.290	.466	74	.454	.507
33	.444	.242	75	.367	.416
34	.517	.306	76	.358	.510
35	.478	.232	77	.260	.208
36	.493	.165	78	.341	.584
37	.185	.354	79	.348	.612
38	.409	.306	80	.583	.285
39	.467	.435	81	.778	-.319
40	.336	.349	82	.292	.569
41	.138	.217	83	.484	.267
42	.587	.420			

TABLE 4

Coefficients for Classrooms and Scales on Characteristic Components of Classroom Judgements

Scale	Component I Class A	Component I Class B	Component I Class C	Component I Class D	Component II Class A	Component II Class B	Component II Class C	Component II Class D	Component III Class A	Component III Class B	Component III Class C	Component III Class D
1	3.32	8.65	-8.25	5.85	2.89	-3.87	11.96	.80	-.25	2.35	10.61	3.00
2	-2.89	3.30	-10.79	6.85	6.13	-2.18	11.26	4.61	-5.31	7.41	10.61	-3.27
3	3.67	6.49	-5.23	5.23	3.17	-3.75	14.00	-.14	1.26	3.71	10.25	3.83
4	5.85	8.89	-5.92	5.54	-2.37	-3.35	7.20	.12	2.42	2.60	14.80	2.97
5	.66	-8.17	-15.65	-1.53	-1.66	-12.25	.38	-2.87	.25	7.38	-.72	1.92
6	5.45	7.40	-4.70	4.46	-1.78	-3.17	8.86	3.43	4.10	3.37	15.07	3.02
7	1.56	2.75	-9.24	2.61	-3.76	-4.24	3.76	.01	5.93	8.66	15.86	.23
8	6.16	2.62	-.34	-2.63	-.11	-.54	1.38	2.40	-.38	3.97	6.59	1.01
9	7.09	-7.22	11.41	1.52	-1.62	-.98	-4.02	-1.84	1.65	-.49	.64	3.51
10	8.24	-8.28	2.58	4.34	-.51	-3.03	6.57	-.73	3.68	3.22	6.60	2.61
11	-6.95	-10.15	-10.28	-5.28	-3.3[illegible]	-12.10	-17.33	-.40	-3.45	-8.13	4.75	-5.46
12	-3.61	-6.40	-9.48	-1.30	8.72	4.86	-14.61	19.33	-6.42	-9.11	3.11	-10.41
13	-9.26	-10.03	-9.59	-7.93	-10.16	-11.88	-18.11	.88	-2.30	-4.25	1.95	-6.09
14	6.07	-4.12	-8.41	5.11	5.26	2.57	-13.44	15.14	-4.79	-2.63	1.42	-6.09
15	-2.84	3.94	4.93	3.44	-1.54	-1.85	-2.60	.45	5.24	-5.82	-14.88	-15.01
16	-14.48	-1.81	-4.59	5.25	9.35	1.50	4.13	-2.70	-3.70	-11.85	-4.98	-2.94
17	4.29	2.96	-2.10	-2.30	-2.17	-3.50	1.68	2.82	1.02	4.26	7.84	-2.09
18	-2.32	4.28	1.38	2.91	2.92	1.05	4.32	3.26	-4.32	5.91	10.34	-1.16
19	3.40	8.17	10.99	10.68	1.83	-2.39	-5.63	-2.70	-.64	-3.77	-.69	-2.25
20	-4.84	6.91	10.33	-.26	5.39	.08	-4.78	.79	-5.03	-10.45	-2.73	-9.39
21	-4.14	-6.35	-6.46	-6.14	2.39	-7.18	3.93	5.12	-10.72	-13.55	-2.29	-10.42

TABLE 5

Observed Orders and J-Scale for First Principal Component

Orders	Frequency	J-Scale I	J-Scale II
ABCD	1		
ADBC	3		
ADCB	1	BDAC	DCBA
BADC	2	DBAC	(CDBA)
BDAC	3	DABC	CBDA
BDCA	1	ADBC DACB	(BCDA) (CBAD)
CADB	1	ADCB	(BCAD)
CBDA	2	(ACDB)	(BACD)
DABC	2	CADB	ABCD
DACB	1		
DBAC	1		
DBCA	1		
DCBA	1		

The Orders in parentheses were not observed.

TABLE 6

Observed Orders and J-Scale for Second Principal Component

Order	Frequency	J-Scale
ABDC	1	
ACBD	1	CDAB
ADBC	1	DCAB
BADC	1	(DACB)
CADB	5	DABC
CDAB	4	ADBC
CDBA	1	ABDC
DABC	5	BADC
DCAB	1	

The order in parentheses was not observed.

TABLE 7

Observed Orders and J-Scale for Third Principal Component

Order	Frequency	J-Scale
ABCD	1	
ACDB	1	CADB
BDAC	1	ACDB
CABD	4	(ADCB)
CADB	2	DACB
CBAD	3	(ABDC)
CBDA	3	(BADC)
CDAB	1	BDAC
CDBA	3	
DACB	2	

The orders in parentheses were not observed.