AUTHOR	Nelsen, Edward A.; Ray, William J.
TITLE	Observational Ratings of Teaching Performance: Dimensionality and Stability.
PUB DATE	Aug 83
NOTE	23p.; Paper presented at the Annual Meeting of the American Psychological Association (Los Angeles, CA, August 26-30, 1983).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Beginning Teachers; Correlation; Data Collection; Elementary School Teachers; Elementary Secondary Education; Factor Analysis; *Interrater Reliability; *Observation; *Rating Scales; Secondary School Teachers; *Teacher Evaluation; Test Construction
IDENTIFIERS	*Teacher and Pupil Performance Ratings; *Unidimensionality (Tests)

ABSTRACT

The investigation examined relationships among scales for observing and rating teacher performance. Beginning teachers with varying levels of professional experience (2, 9, and 16 months) were rated by pairs of observers on two occasions. Intercorrelations across occasions fell between .5 and .8. Interrater agreement ranged between .5 and .9. Factor analyses revealed about 67 percent common variance among the scales. Two rotated factors characterized "direct instruction" and "classroom control" dimensions. The extent of unidimensional variance is discussed in relation to underlying "true" versus "attributional" (halo effects) sources of common variance. (Author)

Observational Ratings of Teaching Performance: Dimensionality and Stability

Edward A. Nelsen and William J. Ray

College of Education

Arizona State University

For copies or additional information write to:

I. D. Payne Laboratory
College of Education
Arizona State University
Tempe, AZ 85287

This paper was prepared for the Annual Meeting of the American Psychological Association, Los Angeles, CA, August 26-30, 1983.

The study was conducted under the auspices of the Office of Professional Field Experiences, College of Education, Arizona State University.

Observational Ratings of Teaching Performance: Dimensionality and Stability

The investigation examined relationships among scales for observing and rating teacher performance. Beginning teachers with varying levels of professional experience (2, 9, and 16 months) were rated by pairs of observers on two occasions. Intercorrelations across occasions fell between .5 and .8. Interrater agreement ranged between .5 and .9. Factor analyses revealed about 67% common variance among the scales. Two rotated factors characterized "direct instruction" and "classroom control" dimensions. The extent of unidimensional variance is discussed in relation to underlying "true" versus "attributional" (halo effects) sources of common variance.

Observational Ratings of Teaching Performance:

Dimensionality and Stability

Edward A. Nelsen and William J. Ray

Arizona State University

INTRODUCTION

The investigation concerns consistency among observers' ratings of teaching performance. Three forms of consistency are at issue: (1) cross-rater agreement--do persons who simultaneously observe teachers and pupils agree with one another? (2) cross-occasion stability--are ratings of the same teacher across occasions similar? (3) dimensional consistency--are different aspects of teaching performance rated similarly? Interrater agreement, stability, and dimensionality, are elements that are integral for analyses of the generalizability for any set of observations (Shavelson and Dempsey-Atwood, 1976; Shavelson and Webb, 1981).

Specifically, the report describes a study of relationships among 16 scales for observing and rating teaching performance. The rating scales comprise the Teacher and Pupil Performance Ratings (TePPR), a new instrument for assessing teaching performance, including aspects of pupil behavior and classroom environment that reflect teaching effectiveness. (Nelsen, Ray, Knight, and Brook, note 1). This report presents data concerning interrater agreement among the observers and concerning the stability of ratings on each scale, as the same teachers were rated on two occasions. The report also describes the extent to which the 16 scales intercorrelated with one another, that is, the proportion of variance among the scales that was common, and the factorial structure of the scales.

Background

Two decades ago, in the first Handbook of Research Teaching, Medley & Mitzel (1963) declared that rating approaches had proven "uniformly unsuccessful in yielding measures of teaching skill." A major source of unreliability and invalidity of ratings, the authors noted, was contamination of measures by halo effects, i.e., the influence of raters' general impressions upon their specific judgments across items on the instrument. They further pointed out that halo effects spuriously inflate (a) coefficients of observer agreement, (b) stability coefficients, and (c) internal consistency among items on a scale.

A more tempered appraisal of the utility of observational ratings was presented by Rosenshine & Furst (1973) in the Second Handbook of Research on Teaching. Based upon earlier reviews of studies in which both rating and category systems were used to predict student achievement, (Rosenshine & Furst, 1971; Rosenshine, 1971) they concluded that the most significant results had been obtained using rating scales, although certainly not all rating scales predicted student learning. An advantage of rating scales, they noted, is the possibility for the observer to process many cues before making a decision. A disadvantage, on the other hand, is that specific details about the sequence, context, and forms of teacher behavior are typically not provided by rating methods.

Rating scales, and other measurement procedures that rely upon perceptions and attributions by observers, yield data that are contaminated by observer errors. Such errors include halo effects and other expectancy effects, differential interpretations of key terms, and judgments that vary because different standards of comparison are employed by different raters. (Cooper, 1981; Fiske, 1978). Measures most vulnerable to such observer errors are those in which key terms and instructions are ill-defined and vague. For example, many teacher rating scales elicit judgments about general characteristics, such as warmth, enthusiasm, or sense of humor, while failing to specify the referent behaviors upon which the observer should focus. Also, rating scales often elicit judgments about characteristics without specifying the situational context, temporal boundaries, or other essential facets that might focus the observers' attention upon specific events (Fiske, 1978).

Critics of ratings scales and attributional measures advocate observational procedures which focus upon specific, narrowly defined acts that can be reliably coded (Medley & Mitzel, 1963; Fiske, 1978). The development and use of such procedures, which have been characterized as "low inference" measures (Rosenshine & Furst, 1973), have undoubtedly contributed to the description and analysis of teaching and learning processes (cf. Good & Brophy, 1978). Evidence concerning particular teacher and pupil behaviors that are indicators of instructional effectiveness has been accumulating, but, to date, no set of specific behavioral indices has emerged as sufficiently basic, comprehensive, or consensually accepted that it could serve as an indicator of competency or general teaching performance (Rosenshine & Furst, 1973).

If low inference measures cannot satisfy the need for economical and comprehensive performance appraisals, and if rating procedures continue to be used despite their unreliability, then evaluators should concentrate upon improvement of rating instruments and reduction of observer errors.

A variety of methods has been employed to reduce halo and increase the accuracy of ratings (Cooper, 1981). Cooper's review of these studies suggested that four methods were most promising as means of reducing illusory halo: increasing rater-ratee familiarity, using multiple raters, rating from current exposure, and obtaining ratings of central, irrelevant categories. Cooper also noted the need for more basic research on how perceptual processes affect rater error.

Meanwhile, the demand for comprehensive indicators of teaching performance and competency continues to grow, as policies and procedures are being developed for certification of competency, tenure decisions, and merit pay. Despite their flaws, rating procedures have continued to serve for these functions and new scales have continued to be developed. For example, the states of Georgia and South Carolina have invested substantial sums of money developing instruments and procedures to certify beginning teachers (Capie, W., Johnson, C. F., Anderson, S. J., Ellet, C. D., & Okey, J. R., note 2; Stulac II, J. F., Gettone, V. G., and others, note 3).

The Teacher Performance Assessment Instruments (TPAI; Capie et al, note 2) and the Assessments of Performance in Teaching (Stulac et al, note 3), were developed to assess minimum proficiency of beginning teachers. These instruments have incorporated improved methods for observing and judging performance. Observer training programs have been established, and the conditions for observing teachers have been structured and standardized. Ratings are obtained on several occasions by at least two raters, so data can be analyzed to determine the extent to which the ratings are generalizable across occasions and raters (Capie, note 4). However, the instruments were designed for a specific purpose, i.e., to elicit discrete judgments concerning the presence or absence of certain minimum proficiences, rather than to measure a broader range of differences in performance levels. The characteristics to be assessed by the instruments were determined by surveys of teachers' and other professionals' opinions concerning "essential competencies", rather than on the basis of systematic theory or research on characteristics of effective teachers. Furthermore, because of the large number of characteristics encompassed by these instruments, the time and costs for observing each teacher are substantial.

A review of teacher observation instruments reported in Simon & Boyer (1970) and Borich and Madden, (1977) did not yield examples of teacher rating instruments that were satisfactory for brief, but comprehensive observational ratings of teacher performance. That is, there appeared to be no instrument that (a) focused upon aspects of teaching performance and pupil behavior that had been shown by research to be related to teaching effectiveness, (b) specified aspects of performance that represented unsatisfactory, satisfactory, and excellent performance, (c) was sufficiently concise to broadly assess teaching performance in an hour or less, (d) and was, at the same time, sufficiently comprehensive to yield an overall assessment of teaching performance.

The Teacher and Pupil Performance Ratings (TePPR) is a new instrument developed to assess performance of beginning teachers in classrooms. The TePPR was designed to provide a comprehensive but brief appraisal of a teacher's performance in the classroom, including cognitive, affective, and interactional aspects of teaching. The TePPR also assesses aspects of pupil behavior and the classroom environment that presumably relate to instructional effectiveness. Certain of the performance dimensions, i.e., clarity of presentation, pupil engagement, and range of interaction, were derived from studies of characteristics associated with instructional effectiveness (cf. Rosenshine & Furst, 1973; Good & Brophy, 1978; Morliave, note 5). Other aspects of performance, e.g., physical organization of the classroom and demonstration of personal regard, were included to study their potential validity as performance indicators.

The scales were designed to differentiate between levels of performance, ranging from poor or unsatisfactory to excellent, as well as to discriminate between adequate and inadequate performance. The primary purpose for developing the TePPR was to provide descriptive data to account for on-the-job performance of graduates from teacher education programs at Arizona State University. In its current form, and until predictive validity studies have been completed, it is recommended that the instrument be used only for such descriptive or research purposes, and not as part of an assessment tool for decisions about individual teachers.

As part of the development of the TePPR, data on performance levels of teachers with different levels of experience were gathered as evidence of construct validity. Also, data concerning interrater agreement, stability of ratings, and intercorrelations among the scales were obtained. These data provide basic evidence concerning the reliability or generalizability of the observations. This report presents these data. It also presents analyses of the factorial structure and of the extent of unidimensionality (or halo) that is manifested in the ratings.

Method

Sample

Recent graduates from teacher education programs at Arizona State University (ASU) comprised the target population. The study included beginning elementary and secondary teachers who had been employed in seven public school districts within a proximity of about 20 miles of the campus. The schools in which these teachers taught varied widely with respect to demographic characteristics of students. They included suburban, inner city, and semi-rural communities; and lower and middle income neighborhoods. All recent graduates who were employed as teachers in these districts were asked to allow observers to schedule two visits to their classes. All but three teachers agreed.

The sample included three groups of graduates, each with successively greater levels of professional experience, as follows:

Group A consisted of 14 beginning teachers with only one to two months of professional teaching experience. The grade levels they taught ranged from kindergarten to 11th grade.

Group B consisted of 35 teachers with five to eight months of experience. Their grade levels also ranged from kindergarten to 11th grade, including some ungraded classes such as home economics, music, and physical education.

Group C included 14 second year teachers who were observed between their 14th and 18th month of teaching. Their grade levels ranged from kindergarten through 6th grade.

Observers

The observers were faculty members and graduate assistants from the College of Education. Their backgrounds were heterogeneous, but all were familiar with public school activities and procedures, and most had teaching experience. Fifteen observers participated in a four-hour orientation and training program prior to the Spring, 1982 studies. Subsequent reliability checks revealed that six of the eight pairs demonstrated agreement greater than .50 (product moment correlations) on at least 13 of the 16 scales. Two rater pairs revealed substantially poorer agreement, and their observations were excluded from the Group B data base.

Four experienced observers provided on-the-job training to four novice observers for the Group C observations. The interrater agreement levels for all the pairs exceeded the criterion of .50 for 13 of the 16 scales.

The Teacher and Pupil Ratings (TePPR) Scales

The TePPR consists of sixteen scales, twelve which describe teacher behaviors or aspects of performance inferred from behavior; one which characterizes the physical aspects of the classroom environment; two which represent pupil behavior; and one which consists of an overall judgment of teaching performances (Nelsen et al., note 1; see appendix for copy of the instrument). The ratings level for each scale range from (1), representing "poor", to (5), representing "excellent"; (3) represents "adequate" performance. Descriptive adjectives define these varying levels for each scale.

The instructions stipulate that observation periods last 45 - 60 minutes, although experienced observers can complete the task in as little as in 30 minutes under optimal conditions. Instructions also state that ratings should be based only on current performance during the session, i.e., excluding recollections from previous observations or other persons' reports about the teacher. Observers are also instructed to signify "no basis for judgment" if classroom activities did not provide a sufficient basis to observe behavior and form a judgment on a particular scale.

Thus, the TePPR employed the following procedures to reduce observer error: using multiple raters, rating from current exposure, rater training, and behaviorally specific rating scales. These design features were based primarily upon Fiske's (1978) suggested strategies for personality assessment. They also correspond with the strategies for reducing halo suggested by Cooper (1981), although development of the TePPR (Nelsen, et al, note 1) preceded our discovery of the Cooper article.

Procedures

Each teacher was observed simultaneously by the same pair of observers on each of two occasions. Each observation session lasted 30 to 60 minutes. The observations were scheduled within three to five weeks of one another. Principals and teachers were asked to participate in the project by letter. Visits were scheduled in advance via phone calls. Confidentiality of the ratings was assured, in that teachers were told that no one other than project staff could see the ratings. Teachers themselves were not shown their own ratings.

Raters were instructed to compare their ratings following each session. Under no circumstances, however, were ratings to be changed on the basis of these cross-checks.

The three groups were constituted of beginning teachers with varying levels of experience. Group A, teachers with one to two months of experience, were observed in Fall, 1982. Group B, with five to eight months of experience, and Group C, with fourteen to eighteen months of experience were observed in Spring, 1982. Eight of the 21 teachers in Group C had been observed previously, one year earlier, by different observers, employing an earlier version of the instrument.

Results

Interrater Agreement

One basis for evaluating the reliability of the observations is provided by data on interrater agreement. Intercorrelations between the ratings based upon simultaneous observation are presented in Table 1, separately for the first and second occasion. For each scale and each occasion, a set of three figures is presented, representing the agreement coefficient for each of the three groups with differing experience levels. For occasion 1, most of the coefficients fell between .5 and .9. For occasion 2, most fell between .5 and 1.0. The median value for the two occasions were .68 and .76, respectively.

On 13 of the 16 scales the agreement coefficients were at least .50 or greater for at least five of the six reliability studies (within the three experience level groups on the two occasions). The reliability coefficients were slightly below this standard for Scale E, Sensitivity to Pupil Comprehension; Scale K, Range of Teacher Interaction; and Scale L, Classroom Management. Two scales revealed agreement coefficients greater than .70 for all groups on both occasions: Scale F, Adaptation to Individual Differences, and Scale J, Pupil Self Control and Responsibility.

Stability of Ratings

A second purpose of the study was to determine the stability of ratings across occasions. Data describing the stability provide another basis for assessing the reliability of the ratings. Correlations between the ratings on the two occasions are included in Table 1, but for ease of comparison, they are also presented separately in Table 2. Most of the stability coefficients were between .5 and .8. Indeed, for each scale, the stability coefficients for at least two of the three experience groups were .5 or greater, with the slight exception of Scale K, for which the coefficients were .71, .48, and .15. The three stability coefficients were quite consistent across the experience level groups for certain scales (A, H, and I). However, they varied considerably for other scales, especially scales B, G, K, and N. This variability would seem to be attributable in large part to sampling error, i.e., as a result of the small size of the samples, especially of Groups A and C. Therefore, it would probably be unwise to infer trends concerning differences in the stability coefficients. Indeed, there did not seem to be any overall tendency for the stability coefficients to be consistently higher or lower for the more versus less experienced teachers.

Dimensionality

Another primary issue in the investigation concerned the dimensionality of the ratings. Data concerning the dimensionality among the scales were provided with factor analyses of the ratings. Data for Group B only were analyzed, since the Ns for groups A and C were too small to yield stable factors. Using the Statistical Programs for Social Sciences (SPSS) principal components analysis program, data were analyzed separately for occasion 1 and 2.

The extent of unidimensionality among the ratings on all scales is reflected in the percent of variance explained by the first principal component. Percentages of variance accounted in the intercorrelation matrices of the first and second occasions were 66.6 and 68.0, respectively.

There is also evidence that an additional basic dimension may be differentiated within the matrices, reflected in the loadings on the second principal component. Employing the criterion of accepting all principal components with eigenvalues greater than 1.0, the first two components were retained for both occasion 1 and 2. Following Kaiser's (1958) varimax procedure, these components were subjected to orthogonal rotation. The results of these analyses are presented in Table 3.

The factors for both occasions are similar. For both occasions, Factor 1 includes loadings from all scales except Scales J., Pupil Self Control and M., Classroom Control. Although all other scales load on this Factor, among the high loadings that define the factor are: C., Presentation of Subject Matter; E., Sensitivity to Pupil Comprehension; G., Quality of Feedback; K., Range of Teacher Interaction; L., Classroom

Management; and N., Quality of Planning, as well as P., Overall Judgment of Teaching Effectiveness. These scales, as well as the other scales, include aspects of instructional directness including effective planning, and management, interaction with many students, subject matter knowledge, and clarity of presentation.

The second factor, which was similar for both occasions, was most clearly defined by the two scales concerning behavioral control: J., Pupil Self Control; and M., Classroom Control. The loadings of these scales on the factor were greater than .8 on both occasions. This factor also included scales with moderate loadings, i.e., between .40 and .60 on B., Clarity of Assignments and Smoothness of Transitions; H., Demonstration of Personal Regard; I., Pupil engagement; L., Classroom Management; and P., Overall Judgment of Effectiveness.

Discussion

The correlations describing interrater agreement indicate that judges with some knowledge of teaching and minimal training can achieve moderate to high agreement when observing and rating a given classroom session with the TePPR scales.

The interobserver agreement was slightly lower on the first observation session than on the second, i.e., a median of .68 versus .76. The higher agreement for the second occasion may result, at least in part, from the comparisons and communication between the raters that followed the first session. That is, they may have influenced one and/or others' judgments concerning aspects of the teachers' performance, and subsequently remembered these judgments on the second occasion. These communications may have also inflated the stability coefficients, which were also moderate (.5 to .8) for most scales. A design which would eliminate such spurious inflation of the stability coefficients and the second occasion agreement coefficients would be provided by a scheme in which the observations were conducted by different pairs of observers on the two occasions. We recently employed this design in a study in which the teachers were rated on separate occasions by different observers.

The factor analytic results reveal a fairly high degree of unidimensionality among the ratings on the 16 scales. This unidimensionality may emanate from two sources. First, aspects of teaching performance and pupil behaviors that reflect effective instruction presumably are integrated and overlapping. Cooper (1981) refers to such interrelationships as "true halo." Much as the cognitive skills that underlie intellectual adaptation are manifested in an intellectual "g" factor, so do mutually related teaching skills that underlie teaching effectiveness manifest themselves in a "g" factor. Unfortunately, aspects of teaching performance and pupil behaviors that reflect teaching effectiveness may also be confounded in the minds of the observers. Thus, perceptions, inferences, and attribution of skill levels on some, if not all of the scales, may have been contaminated by an underlying evaluative dimension, i.e., which Cooper refers to as "illusory halo effects" among observer judgments. The influence of these illusory halo effects, as well as true halo, are both reflected in the high common variance or unidimensionality among the scales.

To a large degree, the data preclude discrimination between these two sources of unidimensional variance among the scales. A research strategy to disentangle the true halo from the illusory halo is needed. Presumably, a systematic program of research to identify sources of such attributional errors in perception of teachers should include both coding (low inference) and ratings (high inference measures).

Reference Notes

Note 1 Nelsen, E.A., Ray, W.J., Knight, C., & Brook, W. Teacher & Pupil Performance Ratings (TePPR). Tempe, AZ: College of Education, Arizona State University, 1981.

Note 2 Capie, W., Johnson, C.F., Anderson, S.J., Ellet, C.D., & Okey, J.R. Teacher Performance Appraisal Instruments (Rev.) Atlanta, Ga.: Georgia Department of Education, 1980.

Note 3 Stulac II, J.F., Gettone, V.G., Stone, J.W., Worthy, D.H., Maiden, M.L., Stokes, M.G., & Thompson, S.A. Assessments of Performance in Teaching: Field Study Instrument: Columbia, S.C.: South Carolina Educator Improvement Task Force; July, 1981.

Note 4 Capie, W. Sampling considerations in classroom research. Paper presented at annual meeting of the American Educational Research Association, Los Angeles, April, 1981.

Note 5 Morliave, R.S. A review of findings of Phase II. San Francisco, CA: Far West Laboratory of Educational Research & Development, May, 1976. (33 pages) ERIC #ED 157871.

References

Borich, G.D. & Madden, S.K. Evaluating classroom instruction: A source book of instruments. Reading, MA: Addison-Wesley, 1977.

Cooper, W.H. Ubiquitous halo. Psychological Review, 1981, Vol. 90, No. 2, 218-244.

Fiske, D.W. Strategies for personality research. San Francisco: Jossey-Bass, 1978.

Good, T. & Brophy, J. Looking in classroom. (2nd ed.). New York: Harper & Row, 1978.

Kaiser, H.F. Varimax criterion for analytic rotation in factor analysis. Psychometrica, 1958, 23 , 187-200.

Medley, D.M., & Mitzel, H.E. Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. pp. 247-328.

Rosenshine, B. Teaching behaviors and student achievement. Winsor, Berkshire, England: National Foundation for Educational Research in England and Wales, 1971.

Rosenshine & Furst, N.F. Research on teacher performance criteria. In B.O. Smith (Ed.), Research in teacher education: Symposium. Englewood Cliffs, NJ: Prentice-Hall, 1971, pp 37-72.

Shavelson, R., & Dempsey-Atwood, N. Generalizability of measures of teaching behavior. Review of Educational Research, 1976, Vol. 46, Fall, pp. 553-611.

Shavelson, R.J., & Webb, N.M. Generalizability theory: 1972-1980. British Journal of Mathematical and Statistical Psychology, Vol. 34, 1981, pp. 133-166.

Simon, A. & Boyer, E.G., (eds.). Mirrors for behavior: An anthology of observation instruments. Philadelphia: Research for Better Schools, 1970.

Table 1

Means, Standard Deviations, Inter rater Reliability Coefficients, and Stability Coefficients for TePPR Ratings of Teacher Performance on Two Occasions

		Occasion 1					Occasion 2					
	Group[a]	N_A[b]	N_B	M	SD	r_{AB}	N_A	N_B	M	SD	r_{AB}	r_{12}[c]
A. Organization of Classroom	A	14	13	3.55	.68	.75	12	6	3.16	.80	.76	.60
	B	34	34	3.81	.91	.64	33	35	3.88	.76	.62	.61
	C	21	21	4.09	.68	.45	21	21	4.49	.67	.69	.63
B. Clarity of Assignments: Transitions	A	14	13	3.44	.64	.52	11	6	3.09	1.00	.91	.15
	B	34	33	3.70	.82	.57	31	33	3.57	.99	.83	.53
	C	19	20	3.95	.79	.68	21	20	4.20	.85	.51	.61
C. Presentation of Subject Matter	A	12	11	3.42	.77	.68	10	5	3.30	1.24	.97	.73
	B	31	32	3.76	.95	.75	29	31	3.66	1.15	.77	.73
	C	20	21	3.78	.73	.51	20	17	4.05	.74	.67	.39
D. Questioning	A	12	11	3.61	.80	.56	11	5	3.17	1.10	.99	.50
	B	24	26	3.72	.93	.92	22	26	3.73	.96	.81	.75
	C	20	20	3.75	.63	.56	18	17	3.88	.80	.57	.71
E. Sensitivity to Pupil Comprehension	A	11	12	3.36	.70	.58	12	5	3.55	.93	.87	.64
	B	35	35	3.67	.95	.50	32	34	3.68	.91	.60	.41
	C	21	21	3.73	.66	.10	20	19	4.08	.62	.29	.53
F. Adaptation to Individual Differences	A	12	10	3.59	.75	.82	11	6	3.14	.96	.94	.50
	B	33	35	3.36	.83	.76	28	31	3.50	.98	.62	.68
	C	18	20	3.70	.72	.75	17	19	3.73	.70	.81	.58
G. Quality of Feedback	A	14	13	3.58	1.07	.65	12	5	3.50	.91	.81	.79
	B	33	32	3.81	.90	.75	29	32	3.87	.95	.68	.78
	C	21	21	3.81	.74	.65	20	21	4.31	.65	.40	.33
H. Demonstration of Regard	A	13	13	3.53	.86	.86	11	5	3.56	1.07	.95	.62
	B	34	34	3.98	.93	.84	33	35	3.87	1.01	.65	.73
	C	20	21	3.58	.74	.17	20	20	3.97	.66	.63	.67
I. Pupil Engagement	A	14	13	3.96	.85	.64	12	6	3.58	1.26	.80	.67
	B	35	34	4.16	.87	.81	33	35	4.09	.83	.61	.63
	C	21	21	4.47	.64	.30	21	22	4.53	.63	.77	.60
J. Pupil Self Control, Responsibility	A	14	13	3.66	.70	.75	12	6	3.33	1.08	.75	.75
	B	35	35	4.08	.81	.87	33	35	4.01	.86	.77	.39
	C	21	21	4.02	.94	.76	21	21	4.02	.96	.85	.77
K. Range of Teacher Interaction	A	14	12	3.73	.61	.38	12	6	3.33	.96	.63	.71
	B	35	34	3.63	.97	.85	33	35	3.71	.96	.59	.48
	C	21	21	3.76	.76	.67	20	20	4.30	.71	.17	.15
L. Classroom Management	A	14	13	3.25	.65	.39	12	6	3.04	.86	.86	.70
	B	35	34	3.85	.98	.83	32	35	3.72	1.00	.81	.53
	C	21	21	4.02	.84	.42	21	21	4.18	.89	.66	.68
M. Classroom Control	A	14	13	3.58	.74	.50	12	6	3.29	1.20	.87	.72
	B	35	35	4.24	.88	.79	33	35	4.01	.98	.80	.49
	C	20	21	4.12	.84	.72	21	21	4.26	1.07	.90	.81
N. Quality of Planning	A	13	11	3.41	.72	.76	12	6	3.20	.90	.74	.83
	B	34	35	3.59	.93	.84	33	35	3.35	1.06	.79	.54
	C	19	20	3.72	.78	.46	21	21	3.88	.85	.57	.19
O. Knowledge of Subject Matter	A	12	11	3.35	.72	.83	12	6	3.49	.79	.73	.64
	B	34	32	3.74	.88	.86	33	35	3.61	.90	.76	.69
	C	20	21	3.14	.63	.69	19	21	3.80	.63	.53	.49
P. Overall Teaching Performance	A	14	13	3.52	.81	.78	11	6	3.09	1.13	.84	.56
	B	35	35	3.57	.87	.68	35	35	3.59	1.06	.91	.72
	C	18	19	3.75	.68	.70	20	18	3.89	.93	.86	.71

[a] Group A - Beginning teachers with one to two months of experience
B - Five to eight months of experience
C - Fourteen to eighteen months of experience

[b] "A" and "B" refer to arbitrary designations of each member of the rater pair

[c] Correlation between combined ratings of observer A and B on Occasion 1 with combined ratings of A and B on Occasion 2

Table 2

Stability Coefficients for TePPR Ratings of Teacher Performance on Two Occasions[a]

	Experience Level		
	First two months	Second semester	Second year - spring
	r_{12}[b] N = 11-12	r_{12}[b] N = 26-35	r_{12}[b] N = 18-21
A. Organization of Classroom	.60	.61	.63
B. Clarity of Assignments; Transitions	.15	.53	.61
C. Presentation of Subject Matter	.73	.73	.39
D. Questioning	.50	.75	.71
E. Sensitivity to Pupil Comprehension	.64	.41	.53
F. Adaptation to Individual Differences	.50	.68	.58
G. Quality of Feedback	.79	.78	.33
H. Demonstration of Regard	.62	.73	.67
I. Pupil Engagement	.67	.63	.60
J. Pupil Self Control, Responsibility	.75	.39	.77
K. Range of Teacher Interaction	.71	.48	.15
L. Classroom Management	.70	.53	.68
M. Classroom Control	.72	.49	.81
N. Quality of Planning	.83	.54	.19
O. Knowledge of Subject Matter	.64	.69	.49
P. Overall Teaching Performance	.56	.72	.71

[a] The two occasions were separated by about two to six weeks.

[b] Correlation between combined ratings of observer A and B on Occasion 1 with combined ratings of A and B on Occasion 2.

Table 3

Factor Analysis of Ratings for Two Occasions

	Scale	Occasion 1		Occasion 2	
		Instruction	Control	Instruction	Control
A.	Physical Organization of Classroom	.77*	.23	.61*	.16
B.	Clarity of Assignments/ Transitions	.55*	.57*	.73*	.52*
C.	Presentation of Subject Matter	.74*	.44*	.87*	.28
D.	Effectiveness of Questions	.64*	.62*	.85*	.22
E.	Sensitivity to Pupil Comprehension	.78*	.24	.76*	.51*
F.	Adaptation to Individual Differences	.78*	.28	.68*	.47*
G.	Quality of Feedback	.91*	.22	.73*	.35
H.	Demonstration of Personal Regard	.70*	.41*	.64*	.45*
I.	Pupil Engagement in Tasks	.60*	.64*	.60*	.45*
J.	Pupil Self Control	.07	.98*	.19	.89*
K.	Range of Teacher Interaction	.75*	.12	.72*	.25
L.	Classroom Management	.71*	.51*	.73*	.53*
M.	Classroom Control	.34	.81*	.27	.83*
N.	Quality of Planning	.74*	.40 *	.79*	.38
O.	Knowledge of Subject Matter	.68*	.30	.91*	.19
P.	Overall Judgment	.80*	.49*	.76*	.57*

*loading = > .40

TEACHER AND PUPIL PERFORMANCE RATINGS

(TePPR)

BACKGROUND

NAME OF PERSON BEING OBSERVED	DATE OF OBSERVATION	TIME BEGUN
PROGRAM □ Field Based; □ Campus Based	SCHOOL	DEPARTMENT (IN COLLEGE OF EDUCATION) □ Elementary; □ Secondary; □ Special
GRADE LEVEL/SUBJECT	NAME OF OBSERVER	YEAR OF TEACHING □ First; □ Second; □ Third or more

SETTING (Describe the classroom setting and circumstances present during observation period)

PHYSICAL DESIGN OF CLASSROOM (CHECK ONE OR MORE)

□ Self-contained; □ Open; □ Team teaching; □ Resource room; □ Media center; (OTHER) □ ____________

STAFF PRESENT (SPECIFY IF MORE THAN ONE)

□ Aide(s); □ Co-teacher(s); □ Student teacher(s); □ Others (observers, parents, etc.)

ORGANIZATION OF INSTRUCTION (CHECK ONE OR MORE)

□ Whole class; □ One small group/ individual seatwork; □ Small groups; □ Individualized

INSTRUCTIONAL MODE(S) (CHECK ONE OR MORE)

□ Lecture; □ Question answer; □ Demonstration; □ Individual seatwork; □ Learning centers; (OTHER) □ ____________

SUBJECT MATTER TAUGHT (DURING OBSERVATION PERIOD. ESTIMATE NUMBER OF MINUTES FOR EACH).

MIN.	SUBJECT	MIN.	SUBJECT	MIN.	SUBJECT	MIN.	SUBJECT	MIN.	SUBJECT	MIN.	SUBJECT	NUMBER OF STUDENTS PRESENT
	Reading		Language Arts		Math-ematics		Social Studies		Science			

Comments (distinctive features of the situation, e.g., minority students, gifted class, handicapped students, unusual case, etc.):

PERFORMANCE RATINGS

The instrument is designed to summarize observations and judgments of a teacher's instructional performance and pupils' behavior in a classroom setting. The observation period should spar. approximately 45 to 60 minutes. The ratings should be based on direct observations of the teacher's and pupils' behaviors during one observation period. Information from previous observations, other persons' reports of the teacher's performance, etc., should not influence the ratings of performance on this occasion. **Do not rate performance on a scale if your observation period did not provide you with an opportunity to observe the behavior specified on that scale.**

Basis for judgment. Lessons, activities, and tcacher roles vary from one class period to another. Your opportunity to observe certain types of teacher or pupil behavior will also vary from one class period to another. The "basis for judgment" ratings allow you to indicate whether you had sufficient or insufficient opportunities to observe each type of behavior considered. Check "no basis for judgment" if the lesson did not present any situations in which this form of performance could be observed and evaluated. Check "substantial basis for judgment" if the lesson presented a sufficient number of episodes as a basis for judging performance on this dimension, or if the lesson included situations that prompted or called for observable behavior relevant to this dimension. Indicate "limited" or "moderate" for class periods that provided bases between "no basis" and "substantial."

Developed by: Edward A. Nelsen
William J. Ray
Catharine C. Knight
Weston L. Brook

A. Physical organization of classroom and instructional materials; utilization of space — furnishings efficiently arranged, pupils visible to teacher and vice versa, adequacy of space for small group work; posted rules and directions are visible and readable; work materials are accessible.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Poorly organized, poor visibility, limited accessibility		Adequately organized		Well organized, facilitative	

STRENGTHS/LIMITATIONS

B. Clarity of assignments and smoothness of transitions to instructional activities — preciseness of directions and task structure: promptness of class response.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Unclear directions, confusion, delays		Adequate	Clear directions, smooth, efficient transitions, pupils respond to directions and begin assignments promptly		

STRENGTHS/LIMITATIONS

C. Skillfulness in presentation of subject matter — clarity, relevance of content, comprehensibility of explanations, use of examples.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Vague, confused, stereotypic, fragmented, oversimplified, boring to pupils		Adequate		Clear, precise, complete, coherent, logical, interesting to pupils	

STRENGTHS/LIMITATIONS

D. Effectiveness, frequency, and level of questions — variety (e.g., open and closed questions), relevance, clarity of questions; extent to which questions require student to mentally manipulate information or support an answer with logically measured evidence ("high" level or divergent versus "low" level questions).

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Vague, narrow, stereotyped, unanswerable, or low cognitive questions		Occasional, fairly effective questions		Frequent, clear, varied, answerable stimulating, high cognitive questions	

STRENGTHS/LIMITATIONS

E. Sensitivity to pupil comprehension — responsiveness to pupil confusion, misunderstanding, boredom, distraction.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Insensitive, unresponsive to confusion		Adequate awareness and sensitivity		Sensitive, aware, responsive to pupil understanding	

STRENGTHS/LIMITATIONS

F. Adaptation to individual ability differences of pupils — difficulty of assignments/lessons suitable for ability levels of all pupils; adequate wait-time; activities are challenging to pupils of different ability levels; appropriate pacing.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Instruction too difficult (or easy) for many students or too slow		Difficulty level and pace usually appropriate to most students		Highly responsive and sensitive to all ability levels, appropriate pace	

STRENGTHS/LIMITATIONS

G. Quality of feedback — indication of correct/incorrect pupil responses. Identification and clarification of correct and incorrect elements of the pupil responses (re: performance on worksheets, homework, recitation, etc.)

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Disparaging, vague, or entirely lacking		Adequate		Informative, prompt, clear, helpful	

STRENGTHS/LIMITATIONS

H. Demonstration of personal regard — compliments when appropriate, provides encouragement, courteous, friendly, enthusiastic. Includes verbal and non-verbal reinforcement.

□ 1	□ 2	□ 3	□ 4	□ 5	BASIS FOR JUDGMENT □ No basis; □ Limited; □ Moderate; □ Substantial
Negative, indifferent, vague, disparaging		Moderately effective		Enthusiastic, positive, encouraging	

STRENGTHS/LIMITATIONS

I. **Pupil engagement in tasks** — responsiveness to tasks, attentiveness, and persistence. (Observe at least three times during the class period).

□ 1	□ 2	□ 3	□ 4	□ 5
Low student involvement, less than 25% of pupils engaged or attentive to tasks/activities		Moderate involvement, about 50% of pupils engaged	High involvement, more than 75% of pupils engaged and attentive to tasks/activities most of time	

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

J. **Pupil self control, responsibility for behavior** — pupil compliance with classroom procedures and rules on own volition.

□ 1	□ 2	□ 3	□ 4	□ 5
Pupils act disruptively, require continual monitoring and discipline		Majority of pupils control selves most of time but several do not comply with procedures		Pupils maintain order without direct teacher intervention

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

K. **Range of teacher interaction** — teacher interacts with all pupils, not just a few select individuals or groups, e.g., on basis of ability level or location in the classroom, sex or ethnicity.

□ 1	□ 2	□ 3	□ 4	□ 5
Consistently ignores or criticizes certain children, narrow action zone		Adequate consideration and distribution of attention	Impartially attentive and responsive to all pupils; action includes entire class or group	

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

L. **Classroom management** — appropriate activities, efficient use of time, organization of activities, alternative tasks available for children who complete tasks.

□ 1	□ 2	□ 3	□ 4	□ 5
No activities for some children, poor use of time		Adequate activities and use of time		Appropriate activities provided; efficient use of time

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

M. **Classroom control** — anticipation and control over potentially disruptive situations and behaviors; consistent enforcement of rules, orderly classroom procedures.

□ 1	□ 2	□ 3	□ 4	□ 5
Lack of control, chaos prevails, erratic enforcement of rules		Occasional disruptions, but sufficient order to conduct instruction	Appropriate control and order maintained, few problems, minor problems resolved without disrupting class	

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

N. **Quality of planning for this lesso. /activity** — inferred from organization, evidence of goals, clarity of objectives, availability of resources.

□ 1	□ 2	□ 3	□ 4	□ 5
Poorly planned, fragmented activities, lacking objectives		Adequate planning		Well planned, organized, clear objectives, lessons maintain interest

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

O. **Teacher's knowledge of subject matter** — correctness of information, clarity of explanations, relevance of examples, flexibility, elaboration.

□ 1	□ 2	□ 3	□ 4	□ 5
Deficient in skills/ knowledge, teaches only from manual		Adequate	Mastery of subject, presents from more than one viewpoint, uses good examples	

BASIS FOR JUDGMENT: □ No basis; □ Limited; □ Moderate; □ Substantial

STRENGTHS/LIMITATIONS

P. **Judgement of overall teaching performance during this observation.**

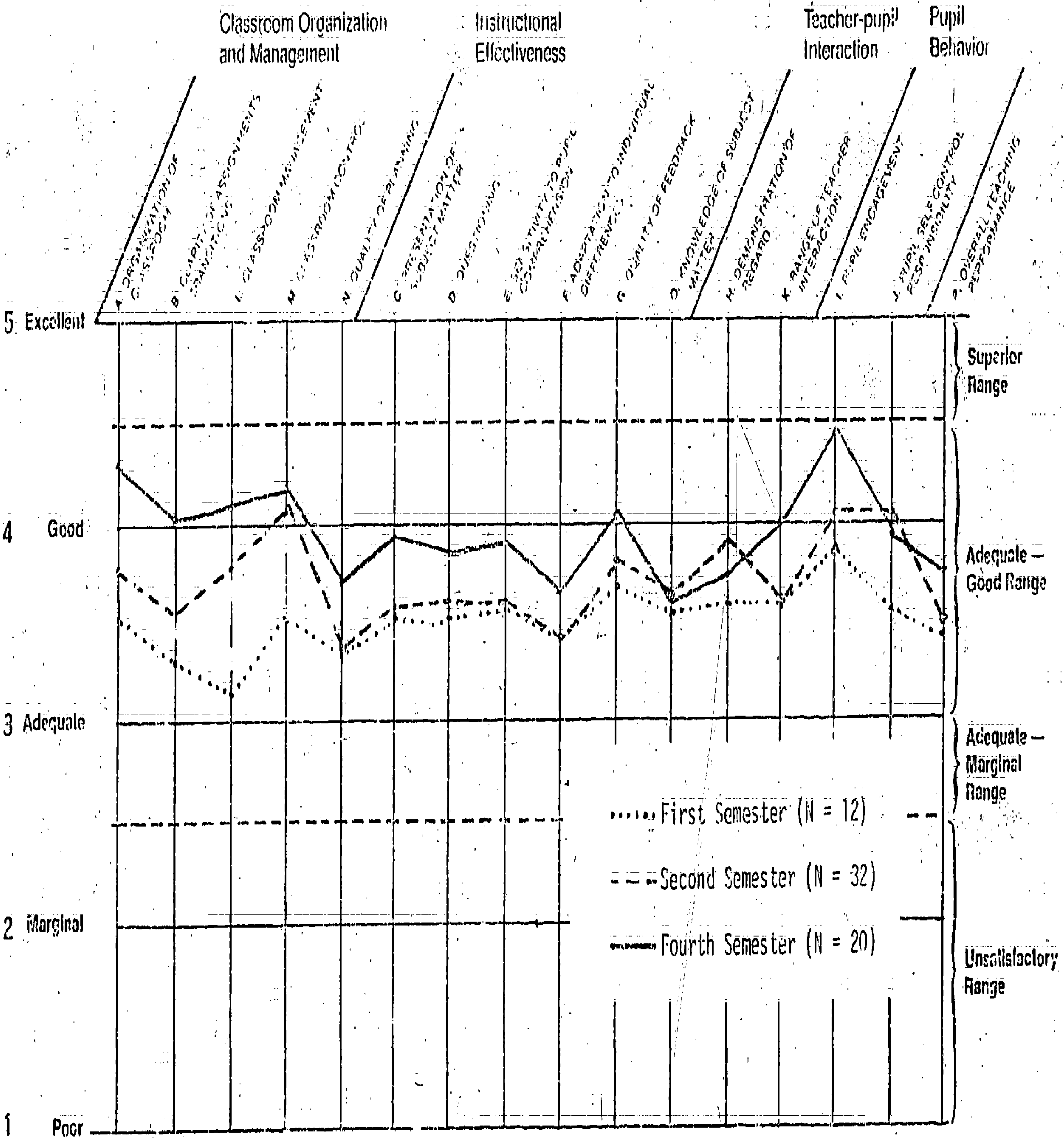
□ 1	□ 2	□ 3	□ 4	□ 5
Not adequate	Marginal	Adequate		Excellent, well planned, stimulating, cohesive session

TIME COMPLETED

(Additional comments on following page).

Comments:

Interview:


Classroom Organization and Management
Instructional Effectiveness
Teacher-pupil Interaction
Pupil Behavior
A. ORGANIZATION OF CLASSROOM
B. CLARITY OF ASSIGNMENTS TRANSITIONS
L. CLASSROOM MANAGEMENT
M. CLASSROOM CONTROL
N. QUALITY OF PLANNING
C. PRESENTATION OF SUBJECT MATTER
D. QUESTIONING
E. SENSITIVITY TO PUPIL COMPREHENSION
F. ADAPTATION TO INDIVIDUAL DIFFERENCES
G. QUALITY OF FEEDBACK
O. KNOWLEDGE OF SUBJECT MATTER
H. DEMONSTRATION OF REGARD
K. RANGE OF TEACHER INTERACTION
I. PUPIL ENGAGEMENT
J. PUPIL SELF CONTROL RESPONSIBILITY
P. OVERALL TEACHING PERFORMANCE
5 Excellent
4 Good
3 Adequate
2 Marginal
1 Poor
Superior Range
Adequate — Good Range
Adequate — Marginal Range
Unsatisfactory Range
First Semester (N = 12)
Second Semester (N = 32)
Fourth Semester (N = 20)