PSYCHOMETRIC PROPERTIES OF CLINICAL PERFORMANCE RATINGS

T. E. DIELMAN
ALAN L. HULL
WAYNE K. DAVIS
University of Michigan

A total of 7931 ratings of 482 third- and fourth-year medical students were gathered over twelve four-week periods. Ratings were made by multiple raters, house officers, and attending faculty, on fifteen behaviorally-anchored rating scales. The data were factor analyzed separately for each of the twelve periods. Two factors consistently emerged, and the congruence coefficients across the twelve periods were high (.72 to .99). The factors were termed "problem solving" and "interpersonal skills" on the basis of their item content. The internal consistency (alpha) coefficients of the scales constructed from the two groups of items and the total fifteen-item scale were high (.83 to .95). Interrater reliability for the individual items ranged from .22 to .37 for attending faculty and from .30 to .51 for house officers. As expected, the interrater reliability was higher for the summed scales than for individual items, ranging from .44 to .61 for house officers and from .36 to .42 for attending faculty.

AUTHORS' NOTE: The authors wish to express their appreciation to Ms. Ann Davidge for her assistance in this study. Inquiries and requests for reprints, and copies of the twelve correlation matrices, principal axis factor matrices, and the Varimax and Promax solutions should be addressed to the senior author, Office of Educational Resources and Research, G1210 Towsley, The University of Michigan Medical School, Ann Arbor, Michigan 48109.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 3 No. 1, March 1980 103-117

Assessment of the performance of medical students during their clerkship is typically a subjective process involving ratings of complex patterns of behavior. Attempts to quantify such ratings have appeared in the medical education literature during the past 25 years.

Diederich (1954) proposed a "critical incidents" technique, derived from studies based on the assessment of Air Force pilots, to assess clinical performance. Cowles and Kubany (1959) adapted Diederich's method and, in interviews with twelve experienced medical faculty members, developed a list of eight characteristics which were considered as most important in preparing for general practice and most easily observed in faculty-student interactions during the third and fourth years of medical school. Those characteristics were identified as: knowledge of medical information; ability to gain and maintain patient's confidence; assumption of responsibility for, and insight into, patient's total problem (medical, social, and emotional); skill in observing, recording, and reporting; skill in developing and verifying hypotheses from patient data; stability under difficult situations; integrity (honesty, ethics, recognition of own limitations); and interest in the profession and in self-improvement. An average of four preceptors rated each of 41 University of Pittsburgh medical students on these eight scales. The intercorrelations among the scales ranged from .14 to .71, with a median correlation of .37. The interrater reliabilities of the average ratings ranged from -.10 (assumption of responsibility for patient's total problem) to .69 (skill in developing and verifying hypotheses), with a median of .49. The interrater reliability of the total of the eight scales was .58.

Cowles (1965) again employed an adaptation of the critical incidents procedure of evaluation and asked medical and nonmedical faculty to sort comments made on 190 University of Pittsburgh medical students into groups. Ten groupings resulted from this process which were identified as: knowledge, rapport with patients, assumption of responsibility, accuracy and thoroughness of observation, diagnostic skill, recognition of limitations, interest/motivation, general intelligence, maturity/poise, and general excellence in medicine. The interrater reliabilities were reported as "high," although no data were presented.

Gough et al. (1964) included a factor analysis of faculty ratings on ten variables in a study of 81 medical students during their third and fourth year clerkships. The variables and their interrater reliabilities for the two years were: clinical judgment (.72, .73); diagnostic skill (.71, .76); mastery of basic sciences (.60, .75); effectiveness of manner with patients (.57, .78); sense of responsibility (.61, .74); independence (.49, .77); identification with medicine (.42, .65); general promise for internship and residency training (.66, .82); and success potential (.69, .80). The intercorrelations among the ten variables ranged from .51 to .93 in the first year of the study and from .61 to .96 during the second year. Three factors resulted from a factor analysis of the first year's data. The first was labeled "medical competence," the second factor was termed "medical identity," and the third factor was termed "medical effectiveness."

In another factor analytic study, Geertsma and Chapman (1967) utilized ratings of 180 medical students on 13 performance scales: overall acceptability as a house officer, total and science GPA, and MCAT scores. The two GPA scores formed one factor (science GPA is a subset of total GPA). Another factor was highly loaded by MCAT verbal and general-information scores, while MCAT science, MCAT quantitative, and general acceptability as a house officer formed separate factors. The first general factor received high loadings from all performance variables, reflecting the tendency of performance ratings to intercorrelate more highly with each other than with grades or MCAT scores. A separate factor was loaded by ratings of ethical standards, likability, and rapport with patients.

On the basis of the above literature review items were selected for the rating instrument for the current study. After discussions with the faculty, departmental committees, and students of The University of Michigan Medical School, a decision was made to develop a set of behaviorally anchored scales to provide specific feedback to the students and to provide the rater with a behavioral (rather than a vague conceptual) definition of the trait under consideration. The purpose of the present study was to provide evidence concerning the factor structure, internal consistency, and interrater reliability of those scales.

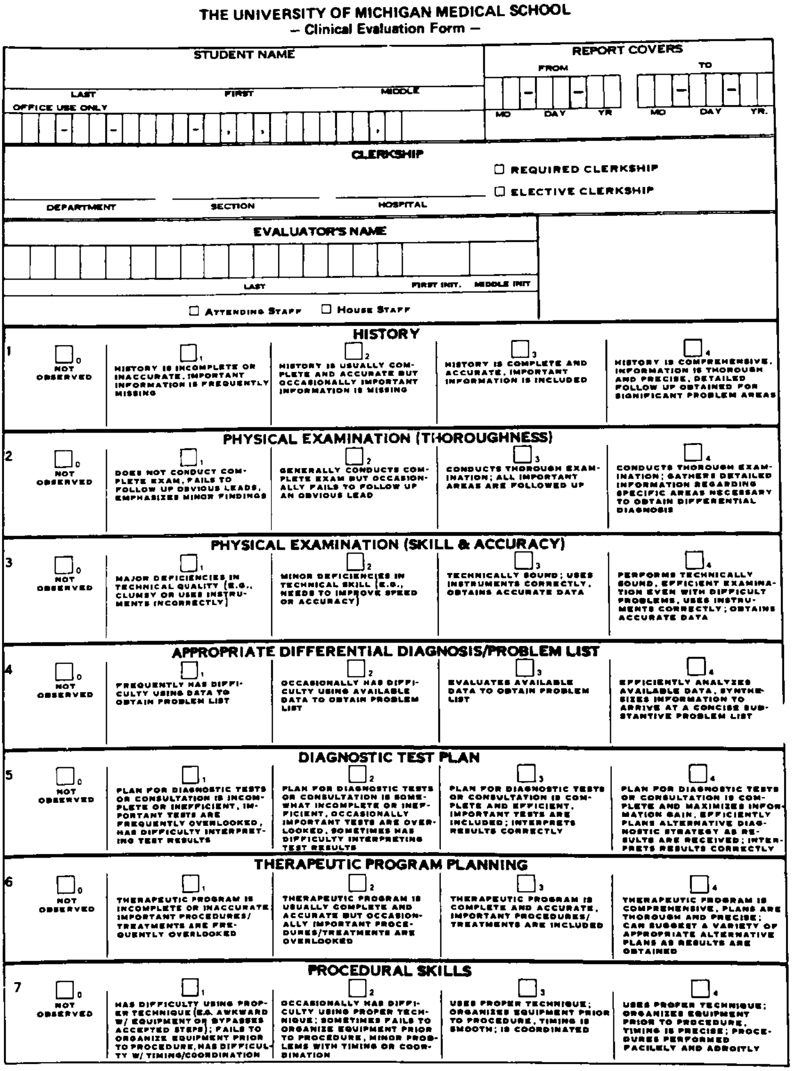
METHOD

DATA COLLECTION

The clinical evaluation form used in the current study consists of fifteen behaviorally based performance scales (see Figure 1). The categories represent a range of clinical and professional skills. The form provides four behaviorally defined categories for rating each of the fifteen performance scales. A space for "not observed" is also provided on each scale. The behaviorally defined categories were developed to provide the students with more specific feedback than would have been provided by adjectives such as "outstanding" or "poor."

The data reported in this study include 7931 ratings gathered over a twelve-month period, from July 1977 to July 1978, on a total of 482 third- and fourth-year medical students. The academic year is divided into twelve four-week periods. Students were rated during each of the periods by two or more house officers and/or attending staff members. This procedure allowed the computation of interrater reliability coefficients for the fifteen performance scales as well as internal consistency (Cronbach alpha) coefficients for additive combinations of those scales.

Separate data bases were constructed for the factor analysis, Cronbach alpha, and interrater reliability computations. Since biases in these analyses may have been the result of an unequal number of evaluations completed for each student, the following sampling procedure was employed in the construction of the



THE UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
– Clinical Evaluation Form –

STUDENT NAME

LAST FIRST MIDDLE

OFFICE USE ONLY

REPORT COVERS

FROM MO DAY YR TO MO DAY YR.

CLERKSHIP

DEPARTMENT SECTION HOSPITAL

□ REQUIRED CLERKSHIP

□ ELECTIVE CLERKSHIP

EVALUATOR'S NAME

LAST FIRST INIT. MIDDLE INIT

□ ATTENDING STAFF □ HOUSE STAFF

HISTORY

1	□ 0 NOT OBSERVED	□ 1 HISTORY IS INCOMPLETE OR INACCURATE, IMPORTANT INFORMATION IS FREQUENTLY MISSING	□ 2 HISTORY IS USUALLY COMPLETE AND ACCURATE BUT OCCASIONALLY IMPORTANT INFORMATION IS MISSING	□ 3 HISTORY IS COMPLETE AND ACCURATE, IMPORTANT INFORMATION IS INCLUDED	□ 4 HISTORY IS COMPREHENSIVE, INFORMATION IS THOROUGH AND PRECISE, DETAILED FOLLOW UP OBTAINED FOR SIGNIFICANT PROBLEM AREAS

PHYSICAL EXAMINATION (THOROUGHNESS)

2	□ 0 NOT OBSERVED	□ 1 DOES NOT CONDUCT COMPLETE EXAM, FAILS TO FOLLOW UP OBVIOUS LEADS, EMPHASIZES MINOR FINDINGS	□ 2 GENERALLY CONDUCTS COMPLETE EXAM BUT OCCASIONALLY FAILS TO FOLLOW UP AN OBVIOUS LEAD	□ 3 CONDUCTS THOROUGH EXAMINATION; ALL IMPORTANT AREAS ARE FOLLOWED UP	□ 4 CONDUCTS THOROUGH EXAMINATION; GATHERS DETAILED INFORMATION REGARDING SPECIFIC AREAS NECESSARY TO OBTAIN DIFFERENTIAL DIAGNOSIS

PHYSICAL EXAMINATION (SKILL & ACCURACY)

3	□ 0 NOT OBSERVED	□ 1 MAJOR DEFICIENCIES IN TECHNICAL QUALITY (E.G., CLUMSY OR USES INSTRUMENTS INCORRECTLY)	□ 2 MINOR DEFICIENCIES IN TECHNICAL SKILL (E.G., NEEDS TO IMPROVE SPEED OR ACCURACY)	□ 3 TECHNICALLY SOUND; USES INSTRUMENTS CORRECTLY, OBTAINS ACCURATE DATA	□ 4 PERFORMS TECHNICALLY SOUND, EFFICIENT EXAMINATION EVEN WITH DIFFICULT PROBLEMS, USES INSTRUMENTS CORRECTLY; OBTAINS ACCURATE DATA

APPROPRIATE DIFFERENTIAL DIAGNOSIS/PROBLEM LIST

4	□ 0 NOT OBSERVED	□ 1 FREQUENTLY HAS DIFFICULTY USING DATA TO OBTAIN PROBLEM LIST	□ 2 OCCASIONALLY HAS DIFFICULTY USING AVAILABLE DATA TO OBTAIN PROBLEM LIST	□ 3 EVALUATES AVAILABLE DATA TO OBTAIN PROBLEM LIST	□ 4 EFFICIENTLY ANALYZES AVAILABLE DATA, SYNTHESIZES INFORMATION TO ARRIVE AT A CONCISE SUBSTANTIVE PROBLEM LIST

DIAGNOSTIC TEST PLAN

5	□ 0 NOT OBSERVED	□ 1 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS INCOMPLETE OR INEFFICIENT, IMPORTANT TESTS ARE FREQUENTLY OVERLOOKED, HAS DIFFICULTY INTERPRETING TEST RESULTS	□ 2 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS SOMEWHAT INCOMPLETE OR INEFFICIENT, OCCASIONALLY IMPORTANT TESTS ARE OVERLOOKED, SOMETIMES HAS DIFFICULTY INTERPRETING TEST RESULTS	□ 3 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS COMPLETE AND EFFICIENT, IMPORTANT TESTS ARE INCLUDED; INTERPRETS RESULTS CORRECTLY	□ 4 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS COMPLETE AND MAXIMIZES INFORMATION GAIN, EFFICIENTLY PLANS ALTERNATIVE DIAGNOSTIC STRATEGY AS RESULTS ARE RECEIVED; INTERPRETS RESULTS CORRECTLY

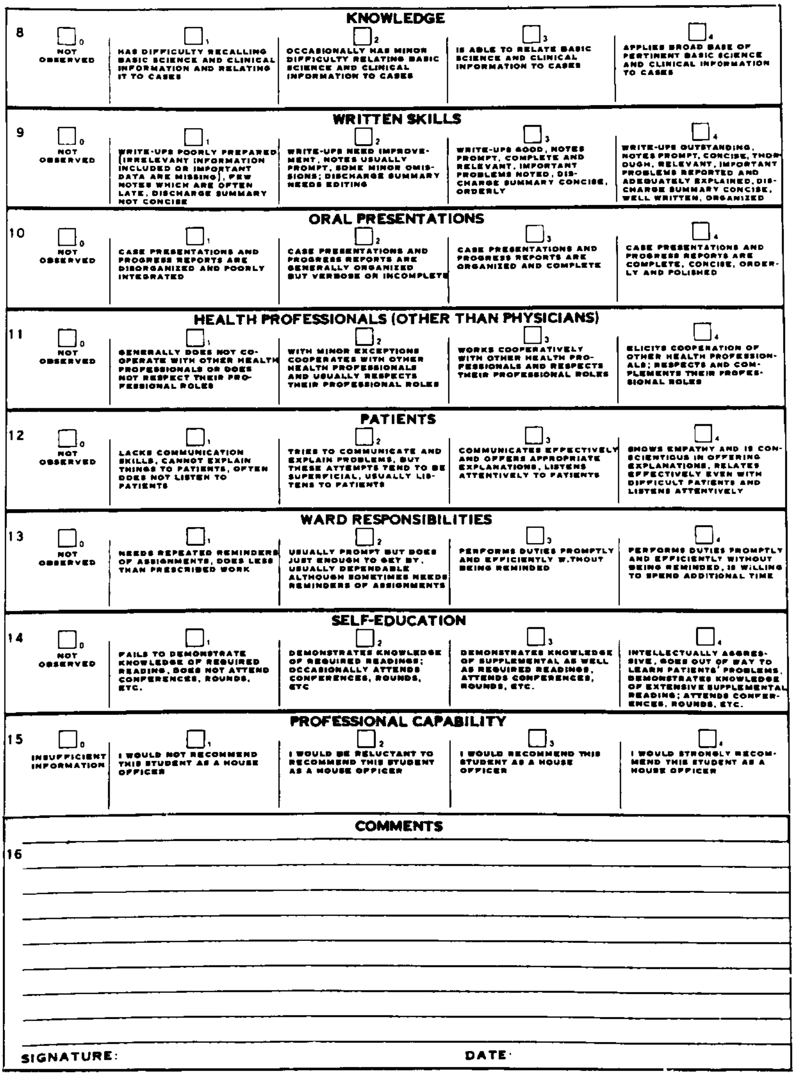
THERAPEUTIC PROGRAM PLANNING

6	□ 0 NOT OBSERVED	□ 1 THERAPEUTIC PROGRAM IS INCOMPLETE OR INACCURATE; IMPORTANT PROCEDURES/TREATMENTS ARE FREQUENTLY OVERLOOKED	□ 2 THERAPEUTIC PROGRAM IS USUALLY COMPLETE AND ACCURATE BUT OCCASIONALLY IMPORTANT PROCEDURES/TREATMENTS ARE OVERLOOKED	□ 3 THERAPEUTIC PROGRAM IS COMPLETE AND ACCURATE, IMPORTANT PROCEDURES/TREATMENTS ARE INCLUDED	□ 4 THERAPEUTIC PROGRAM IS COMPREHENSIVE, PLANS ARE THOROUGH AND PRECISE; CAN SUGGEST A VARIETY OF APPROPRIATE ALTERNATIVE PLANS AS RESULTS ARE OBTAINED

PROCEDURAL SKILLS

7	□ 0 NOT OBSERVED	□ 1 HAS DIFFICULTY USING PROPER TECHNIQUE (E.G. AWKWARD W/ EQUIPMENT OR BYPASSES ACCEPTED STEPS); FAILS TO ORGANIZE EQUIPMENT PRIOR TO PROCEDURE, HAS DIFFICULTY W/ TIMING/COORDINATION	□ 2 OCCASIONALLY HAS DIFFICULTY USING PROPER TECHNIQUE; SOMETIMES FAILS TO ORGANIZE EQUIPMENT PRIOR TO PROCEDURE, MINOR PROBLEMS WITH TIMING OR COORDINATION	□ 3 USES PROPER TECHNIQUE; ORGANIZES EQUIPMENT PRIOR TO PROCEDURE, TIMING IS SMOOTH; IS COORDINATED	□ 4 USES PROPER TECHNIQUE; ORGANIZES EQUIPMENT PRIOR TO PROCEDURE, TIMING IS PRECISE; PROCEDURES PERFORMED FACILELY AND ADROITLY

Figure 1: The Clinical Evaluation Form



KNOWLEDGE

8	□0 NOT OBSERVED	□1 HAS DIFFICULTY RECALLING BASIC SCIENCE AND CLINICAL INFORMATION AND RELATING IT TO CASES	□2 OCCASIONALLY HAS MINOR DIFFICULTY RELATING BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	□3 IS ABLE TO RELATE BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	□4 APPLIES BROAD BASE OF PERTINENT BASIC SCIENCE AND CLINICAL INFORMATION TO CASES

WRITTEN SKILLS

9	□0 NOT OBSERVED	□1 WRITE-UPS POORLY PREPARED (IRRELEVANT INFORMATION INCLUDED OR IMPORTANT DATA ARE MISSING), FEW NOTES WHICH ARE OFTEN LATE, DISCHARGE SUMMARY NOT CONCISE	□2 WRITE-UPS NEED IMPROVEMENT, NOTES USUALLY PROMPT, SOME MINOR OMISSIONS; DISCHARGE SUMMARY NEEDS EDITING	□3 WRITE-UPS GOOD, NOTES PROMPT, COMPLETE AND RELEVANT, IMPORTANT PROBLEMS NOTED, DISCHARGE SUMMARY CONCISE, ORDERLY	□4 WRITE-UPS OUTSTANDING, NOTES PROMPT, CONCISE, THOROUGH, RELEVANT, IMPORTANT PROBLEMS REPORTED AND ADEQUATELY EXPLAINED, DISCHARGE SUMMARY CONCISE, WELL WRITTEN, ORGANIZED

ORAL PRESENTATIONS

10	□0 NOT OBSERVED	□1 CASE PRESENTATIONS AND PROGRESS REPORTS ARE DISORGANIZED AND POORLY INTEGRATED	□2 CASE PRESENTATIONS AND PROGRESS REPORTS ARE GENERALLY ORGANIZED BUT VERBOSE OR INCOMPLETE	□3 CASE PRESENTATIONS AND PROGRESS REPORTS ARE ORGANIZED AND COMPLETE	□4 CASE PRESENTATIONS AND PROGRESS REPORTS ARE COMPLETE, CONCISE, ORDERLY AND POLISHED

HEALTH PROFESSIONALS (OTHER THAN PHYSICIANS)

11	□0 NOT OBSERVED	□1 GENERALLY DOES NOT COOPERATE WITH OTHER HEALTH PROFESSIONALS OR DOES NOT RESPECT THEIR PROFESSIONAL ROLES	□2 WITH MINOR EXCEPTIONS COOPERATES WITH OTHER HEALTH PROFESSIONALS AND USUALLY RESPECTS THEIR PROFESSIONAL ROLES	□3 WORKS COOPERATIVELY WITH OTHER HEALTH PROFESSIONALS AND RESPECTS THEIR PROFESSIONAL ROLES	□4 ELICITS COOPERATION OF OTHER HEALTH PROFESSIONALS; RESPECTS AND COMPLEMENTS THEIR PROFESSIONAL ROLES

PATIENTS

12	□0 NOT OBSERVED	□1 LACKS COMMUNICATION SKILLS, CANNOT EXPLAIN THINGS TO PATIENTS, OFTEN DOES NOT LISTEN TO PATIENTS	□2 TRIES TO COMMUNICATE AND EXPLAIN PROBLEMS, BUT THESE ATTEMPTS TEND TO BE SUPERFICIAL, USUALLY LISTENS TO PATIENTS	□3 COMMUNICATES EFFECTIVELY AND OFFERS APPROPRIATE EXPLANATIONS, LISTENS ATTENTIVELY TO PATIENTS	□4 SHOWS EMPATHY AND IS CONSCIENTIOUS IN OFFERING EXPLANATIONS, RELATES EFFECTIVELY EVEN WITH DIFFICULT PATIENTS AND LISTENS ATTENTIVELY

WARD RESPONSIBILITIES

13	□0 NOT OBSERVED	□1 NEEDS REPEATED REMINDERS OF ASSIGNMENTS, DOES LESS THAN PRESCRIBED WORK	□2 USUALLY PROMPT BUT DOES JUST ENOUGH TO GET BY, USUALLY DEPENDABLE ALTHOUGH SOMETIMES NEEDS REMINDERS OF ASSIGNMENTS	□3 PERFORMS DUTIES PROMPTLY AND EFFICIENTLY W.THOUT BEING REMINDED	□4 PERFORMS DUTIES PROMPTLY AND EFFICIENTLY WITHOUT BEING REMINDED, IS WILLING TO SPEND ADDITIONAL TIME

SELF-EDUCATION

14	□0 NOT OBSERVED	□1 FAILS TO DEMONSTRATE KNOWLEDGE OF REQUIRED READING, DOES NOT ATTEND CONFERENCES, ROUNDS, ETC.	□2 DEMONSTRATES KNOWLEDGE OF REQUIRED READINGS; OCCASIONALLY ATTENDS CONFERENCES, ROUNDS, ETC	□3 DEMONSTRATES KNOWLEDGE OF SUPPLEMENTAL AS WELL AS REQUIRED READINGS, ATTENDS CONFERENCES, ROUNDS, ETC.	□4 INTELLECTUALLY AGGRESSIVE, GOES OUT OF WAY TO LEARN PATIENTS' PROBLEMS, DEMONSTRATES KNOWLEDGE OF EXTENSIVE SUPPLEMENTAL READING; ATTENDS CONFERENCES, ROUNDS, ETC.

PROFESSIONAL CAPABILITY

15	□0 INSUFFICIENT INFORMATION	□1 I WOULD NOT RECOMMEND THIS STUDENT AS A HOUSE OFFICER	□2 I WOULD BE RELUCTANT TO RECOMMEND THIS STUDENT AS A HOUSE OFFICER	□3 I WOULD RECOMMEND THIS STUDENT AS A HOUSE OFFICER	□4 I WOULD STRONGLY RECOMMEND THIS STUDENT AS A HOUSE OFFICER

COMMENTS

16 ______________________________

____

____

____

____

____

____

____

SIGNATURE: DATE·

Figure 1 Continued

data bases for purposes of the internal consistency (Cronbach alpha) and factor analyses:

(1) The only evaluation forms included in the analyses were those on which a house officer or attending staff member rated the student on each of the fifteen scales.
(2) When multiple evaluations of a single student were completed during the same period, at the same location, and by the same evaluator type, one form was randomly selected for that student. This was done to reduce the bias introduced by the possibility that good students were evaluated more or less frequently than poor students.

Based on the above criteria, 1862 completed evaluation forms were included in the Cronbach alpha computations. Since students changed locations in the middle of the medicine rotation, there were 136 cases in which there were two evaluations for a rater type in a period—one for each location. In these cases, one evaluation was selected at random for purposes of the factor analyses. This procedure resulted in a data base of 1726 evaluations for the factor analyses. Twelve replications of the factor analysis were done on this data base, one for each period.

The larger data base for the interrater reliability analysis included 1908 ratings of 389 students by attending staff and 1749 ratings of 355 students by house officers.

DATA ANALYSIS

For each of the twelve periods, correlations were computed and the correlation matrices were factor analyzed by the principal axes procedure. The number of factors was determined by application of the Kaiser "unity rule" (Guertin and Baily, 1970: 115). The twelve factor matrices were rotated by both orthogonal (Varimax) and oblique (Promax) procedures.

The internal consistency of the "factor scores" (simple summation of the salient rating scales on each factor) and the total of fifteen scales were subsequently examined by computing

Cronbach alpha coefficients. The Cronbach alphas were computed separately for house officers and attending staff, both for each of the twelve periods and for a total of all periods.

The interrater reliability for each of the fifteen rating scales was computed separately for house officers and attending staff. Interrater reliabilities for the fifteen scales were calculated by an ordinal data extension of the categorical data computation of intraclass correlation coefficients, as discussed by Landis and Koch (1977).

RESULTS

The principal axes factor analysis resulted in two eigenvalues greater than 1.0 for eight of the twelve periods. On the other four periods, only the first eigenvalue was over 1.0. The first two eigenvalues, percentage of variance accounted for by the first two factors, and the correlation between the two factors resulting from the Promax rotations are shown in Table 1 for each of the twelve periods. Two factors were extracted for rotation for each of the twelve factor analyses, even though the Kaiser criterion indicated only one factor for periods 2, 4, 6, and 9. This procedure was followed in order to check the consistency of the pattern of factor loadings across all periods.

The factor patterns, in fact, were quite consistent, as indicated by the matrix of congruence coefficients presented as Table 2. The lowest coefficients appear for period 9 when compared with all other periods. This was a result of all variables loading highly on the first factor in period 9.

The results of the twelve factor analyses are summarized in Table 3, which presents the median factor loading and the range of the loadings for each of the fifteen variables across the twelve periods. The factor pattern values presented in Table 3 resulted from the Promax rotations.

Factor I has been given the shorthand label "problem solving" for convenience of discussion. It actually includes both procedural skills and cognitive ability. The first factor received its

TABLE 1
Eigenvalues, Percentage of Variance Accounted for by Two Factors, Ns, and Factor Correlations for the 12 Periods

	Eigenvalues		% Variance	N	Factor Correlations (Promax Rotation)
Period	I	II			
1	8.24	1.24	63.2	143	-.68
2	9.60	.68	64.9	154	-.77
3	9.51	1.07	70.5	158	-.69
4	9.84	.59	67.8	158	-.80
5	8.65	1.14	65.3	137	-.68
6	9.33	.70	63.5	153	-.76
7	9.25	1.02	68.4	186	-.73
8	8.57	1.21	65.2	150	-.67
9	8.91	.67	59.1	136	-.79
10	8.12	1.19	62.1	107	-.69
11	9.01	1.12	67.5	134	-.74
12	8.59	1.19	65.2	110	-.68

highest loadings from ratings of ability to use information to arrive at the appropriate differential diagnosis (item 4), diagnostic test planning (item 5), therapeutic program planning (item 6), and knowledge (item 8). Consistently high loadings were also contributed by items 1 through 3—history taking and thoroughness and accuracy in performing physical examinations. In each of the twelve analyses the loadings by these items were substantially higher on Factor I than they were on Factor II.

Four additional items—procedural skills (item 7), written skills (item 9), oral presentations (item 10), and self-education (item 14)—loaded most highly on Factor I in at least nine of twelve analyses. One item, professional capability (item 15), had

TABLE 2
Congruence Coefficients Between Matching Factors Across the 12 Periods

Periods	1	2	3	4	5	6	7	8	9	10	11	12
1	-	.99	.99	.98	.95	.99	.97	.98	.89	.97	.98	.97
2	.97	-	.98	.98	.96	.97	.96	.95	.88	.96	.97	.98
3	.96	.93	-	.97	.95	.99	.97	.98	.88	.98	.96	.98
4	.95	.94	.90	-	.96	.98	.97	.97	.90	.97	.98	.98
5	.83	.86	.81	.88	-	.96	.94	.96	.84	.93	.94	.97
6	.96	.96	.95	.93	.86	-	.98	.99	.90	.98	.96	.99
7	.91	.90	.91	.90	.83	.96	-	.98	.93	.98	.96	.97
8	.91	.92	.93	.89	.84	.95	.94	-	.87	.98	.96	.98
9	.82	.81	.79	.87	.72	.86	.90	.77	-	.93	.90	.91
10	.90	.89	.93	.92	.78	.93	.95	.94	.88	-	.96	.98
11	.94	.93	.90	.97	.85	.90	.90	.85	.88	.88	-	.97
12	.92	.93	.93	.93	.88	.96	.95	.94	.87	.90	.90	-

NOTE: Congruence coefficients for Factor I appear above the diagonal; congruence coefficients for Factor II appear below the diagonal (see Harman, 1960: 257 for the computational formula).

its highest loading on Factor I in seven of the twelve analyses, on Factor II in three of the analyses, and about equal loadings on the two factors in two analyses.

The label attached to Factor II is “interpersonal skills.” This factor received consistently higher loadings from ability to work effectively with health professionals other than physicians (item 11) and relationships with patients (item 12). In all instances except one, ward responsibilities (item 13) showed a higher loading on Factor II than on Factor I.

The internal consistency of the scales defining the two factors was calculated by means of Cronbach alpha coefficients. To calculate the alpha coefficients for the two factors, items 1 through 10, 14, and 15 were included in Factor I, and items 11

TABLE 3
Summary of Factor Pattern Values
(medians and ranges) for the 15 Items over the 12 Periods

	Item	Factor I Problem Solving	Factor II Interpersonal Skill
1.	History	.92 (.57 to 1.08)	.09 (-.12 to .49)
2.	Physical examination (Thoroughness)	.87 (.71 to 1.10)	.18 (-.14 to .37)
3.	Physical examination (Skill and Accuracy)	1.00 (.60 to 1.10)	.00 (-.15 to .49)
4.	Differential diagnosis	1.06 (.89 to 1.15)	-.08 (-.20 to .15)
5.	Diagnostic test plan	1.06 (.93 to 1.22)	-.08 (-.33 to .10)
6.	Therapeutic program	1.06 (.92 to 1.22)	-.09 (-.33 to .11)
7.	Procedural skills	.74 (.46 to .86)	.33 (.18 to .63)
8.	Knowledge	1.02 (.61 to 1.11)	-.03 (-.15 to .45)
9.	Written skills	.83 (.17 to 1.06)	.22 (-.09 to .86)
10.	Oral presentations	.76 (.08 to .96)	.30 (.05 to .94)
11.	Relationships with other professionals	-.14 (-.27 to .06)	1.10 (.95 to 1.18)
12.	Relationships with patients	-.16 (-.28 to .05)	1.11 (.96 to 1.17)
13.	Ward responsibilities	.10 (-.04 to .70)	.92 (.39 to 1.03)
14.	Self education	.71 (.17 to 1.13)	.34 (-.20 to .86)
15.	Professional capability	.65 (.14 to .86)	.44 (.19 to .88)

NOTE: Factor pattern values are equivalent to standardized beta weights in a regression equation, with the observed scores serving as the dependent variables and the factor scores as the independent variables. These values are the correlations of the variables with the factor multiplied by a constant and thus may exceed unity (Harman, 1960: 16-19).

through 13 were included in Factor II. The coefficients were computed separately for the ratings rendered by house officers and attending faculty. These results are presented in the top portion of Table 4. It can be seen that the Cronbach alphas were consistently higher than .8 and that they did not differ substantially for house officers or attending faculty. For both groups, the Cronbach alphas were higher in the case of the problem-solving factor than they were for the interpersonal skills factor, as is expected with a greater number of items.

The interrater reliability coefficients for each item, the problem-solving, interpersonal skills, and total scale are presented separately for house officers and attending faculty in the bottom part of Table 4. The interrater agreement on the fifteen individual items for the attending faculty ranged from .22 to .37, with a median coefficient of .29. The attending faculty interrater agreement for the sum of the fifteen items was .40, while it was .42 for the summation of the twelve items defining the problem-solving factor and .36 for the summation of the three items defining the interpersonal skills factor. For house officers, the interrater agreement for the fifteen single items ranged from .30 to .51, with a median of .36. The house officer agreement was .61 for the summation of the fifteen items, .60 for the summation of the twelve problem-solving factor items, and .44 for the summation of the three interpersonal skills factor items.

DISCUSSION

The consistency of the factor analytic results across the twelve periods indicates that the variables employed in the fifteen-item rating scale reliably cluster into two groups, which have been designated as the problem-solving factor and the interpersonal skills factor. These factors were highly replicable across twelve factor analyses.

The factor which has been referred to in this paper as "problem-solving" resembles a merging of the two factors Gough et al. (1964) chose to term medical competence and medical

TABLE 4
Cronbach Alpha Coefficients and Interrater Reliabilities for the Two Factors and the Total Scale as Rated by House Officers and Attending Faculty (all periods)

	House Officers	Attending Faculty
Cronbach Alphas	(N=767)*	(N=1,095)
Problems solving	.95	.95
Interpersonal skill	.83	.86
Total Scale	.95	.95
Inter-rater Reliabilities	(N=1,749)	(N=1,908)*
Item		
1. History	.34	.34
2. Physical exam (Thoroughness)	.33	.28
3. Physical exam (Skill and Accuracy)	.31	.22
4. Differential diagnosis	.39	.30
5. Diagnostic test plan	.37	.33
6. Therapeutic program	.39	.27
7. Procedural skills	.34	.23
8. Knowledge	.42	.27
9. Written skills	.35	.28
10. Oral presentations	.36	.35
11. Relationships with other professionals	.30	.29
12. Relationships with patients	.30	.31
13. Ward responsibilities	.42	.37
14 Self education	.38	.25
15. Professional capability	.51	.36
Factor I. Problem Solving	.60	.42
Factor II. Interpersonal Skills	.44	.36
Total Scale	.61	.40

*Ns refer to the total number of ratings over the twelve periods.

identity. The interpersonal skills factor in the present study resembles the factor which Gough et al. called medical effectiveness. The results are also similar to the factor analysis conducted by Geertsma and Chapman (1967), who found separate factors formed by performance variables and variables having to do with rapport, likability, and ethical standards. The two factors identified in the current study exhibit a high degree of internal consistency, whether ratings are rendered by attending faculty or by house officers.

The interrater reliability on the single items was somewhat higher for house officers than for attending faculty. One explanation which has been proposed for this result is that house officers spend more time observing the students than do the attending faculty, and thus have more data on which to base their ratings. The strength of interrater agreement, as judged along the admittedly arbitrary benchmarks employed by Landis and Koch (1977), is fair to moderate for both groups of raters. Agreement increases when the individual ratings are summed into factor scores or total scores, and it is recommended that such combined scores be employed for the purpose of summative evaluation. Combining the scales in such a fashion is of little use for formative evaluation purposes, however. When the scales are used for formative purposes in providing student feedback, it is recommended that individual ratings be provided and discrepancies among ratings rendered by different raters be discussed. Another possibility when reporting the results of ratings on individual items is to average ratings made by several observers, thus eliminating the error due to observer variation. This procedure can increase the reliability of ratings on single items considerably, as Printen et al. (1973) have shown. Low interrater reliability seems to be the norm when dealing with independent subjective judgments, and ratings of medical student performance are no exception. Faculty members who are responsible for compiling composite student grades should keep this in mind when evaluating students and strive to collect several independent judgments.

REFERENCES

COWLES, J. T. (1965) "A critical-comments approach to the rating of medical students' clinical performance." J. of Medical Education 40: 188-198.

——— and A. J. KUBANY (1959) "Improving the measurement of clinical performance of medical students." J. of Clinical Psychology 15: 139-142.

DIEDERICH, P. B. (1954) "The 'critical incidents' technique applied to medical education." Research memorandum 54-9, Educational Testing Service.

GEERTSMA, R. H. and J. E. CHAPMAN (1967) "The evaluation of medical students." J. of Medical Education 42: 938-948.

GOUGH, H. G., W. B. HALL, and R. E. HARRIS (1964) "Evaluation of performance in medical training." J. of Medical Education 39: 679-692.

GUERTIN, W. H. and J. P. BAILEY, Jr. (1970) Introduction to Modern Factor Analysis. Ann Arbor, MI: Edwards Brothers.

HARMAN, H. H. (1960) Modern Factor Analysis. Chicago: Univ. of Chicago Press.

LANDIS, J. R. and G. G. KOCH (1977) "The measurement of observer agreement for categorical data." Biometrics 33: 159-174.

PRINTEN, K. J., W. CHAPPELL, and D. R. WHITNEY (1973) "Clinical performance evaluation of junior medical students." J. of Medical Education 48: 343-348.