Medical Education 1987, **21**, 130–137

Dimensionality, internal consistency and interrater reliability of clinical performance ratings

B. R. MAXIM† & T. E. DIELMAN‡

†*Department of Mathematics and Statistics, University of Michigan, Dearborn, Michigan and* ‡*Department of Postgraduate Medicine and Health Professions Education, University of Michigan, Ann Arbor, Michigan*

Summary. A total of 6444 ratings of the financial performance of 424 third- and fourth-year medical students were made by house officers and attending teachers during 12 separate internal medicine rotations. Ratings were based on 13 behaviourally anchored rating scales. One rating was randomly selected per student per evaluator type (house officer and attending teacher) during each of the 12 rotation periods. Ratings were factor analysed separately within each rotation period. Two factors emerged consistently, and congruence coefficients across the 12 occasions were high (0·88 or greater). The factors were labelled 'problem-solving (10 items) and 'interpersonal skills' (three items) on the basis of their content. Internal consistency coefficients of the indices constructed from items in the two factors and the total of the 13 items were high (0·9 or greater) and did not differ substantially when computed separately on the ratings from house officers and attending teachers. Interrater reliabilities on the individual items ranged from 0·14 to 0·33.

Key words: *Clinical competence; *clinical clerkship; internal medicine/educ; psychometrics; Michigan; problem-solving; interpersonal relations

Introduction

The literature on the quantification of medical students' clerkship performance has been reviewed in an earlier article by Dielman *et al.* (1980). In that article the results of psychometric analyses of clinical performance rating scales employed at the University of Michigan Medical School were reported. This rating system has been described in greater detail by Davidge *et al.* (1980). Two factors underlying these ratings were identified: 'problem-solving' and 'interpersonal skills'. These factors were highly replicable across 12 separate analyses (congruence coefficients ranged from 0·72 to 0·99) and exhibited acceptable levels of internal consistency (alpha coefficients ranged from 0·83 to 0·95). The interrater reliabilities of the rating items ranged from 0·22 to 0·37 in the case of ratings made by attending teachers† and from 0·31 to 0·51 in the case of ratings made by house officers. The 'problem-solving' factor identified in the Dielman *et al.* (1980) study was similar in content to a combination of factors which Gough *et al.* (1964) had identified as 'medical competence' and 'medical identity'. The 'interpersonal skills' factor which resulted from the Dielman *et al.* (1980) analysis resembled the factor which Gough *et al.* (1964) had termed 'medical effectiveness'. The Dielman *et al.* (1980) results were also similar to those of Geertsma & Chapman (1967), who identified separate factors formed by performance variables and variables concerning rapport with patients, likeability and ethical standards.

Correspondence: Dr B. R. Maxim, Department of Mathematics and Statistics, 1177 University Mall, University of Michigan, Dearborn, Michigan 48128, USA.

†'Attending' is not an international term. Approximately equivalent to the UK Consultant, it refers (USA) to anybody appointed to the honorary or paid staff of a hospital with the right to admit and treat patients.

Since the Dielman *et al.* (1980) report the clinical performance rating scale has been revised to include 13 rather than 15 items, with behavioural anchors given only at the extremes of the scales, rather than at each scale point. The purpose of the current study was to replicate the 1980 study with the revised scale to determine whether the deletion of the intermediate behavioural anchors and/or reduction in the number of items altered the psychometric properties of the scale. The psychometric properties of interest were the dimensionality of the scale (as determined by several factor analytic replications), the internal consistency (Cronbach alpha coefficients) of the indices formed on the basis of the factor analytic results, and the interrater reliability on the items and the indices. Of secondary interest was a comparison of the psychometric properties of the scale when ratings were made by attending teachers and house officers.

Data collection

The clinical evaluation form used in the current study consists of 13 behaviourally based performance scales (see Fig. 1). Each of these scales represents a range of clinical and professional skills. The form provides six points for rating students on each of the scales, with behaviourally defined anchor points. A space for 'not observed' is provided on each scale.

This form is the first major revision of an evaluation form which was developed in 1977. The purpose of this form (and the original) was to provide the student with more specific feedback through the use of behaviourally defined categories than would have been provided by adjectives such as 'poor' or 'outstanding'. The revised form has been used since July 1980.

The data reported in this study were taken from 6444 ratings gathered over a 12-month period from July 1981 to June 1982 on a total of 424 third- and fourth-year medical students. The academic year is divided into 12 4-week periods. Students were rated by two or more house officers, attending staff and/or other health professionals. This procedure allowed for the computation of interrater reliability coefficients for the 13 performance scales, as well as internal consistency coefficients for additive combinations of these scales.

One data set was used for all analyses. The sampling procedure employed in constructing the data set was as follows:

(1) The only evaluation forms included in the analyses were those completed by house officers and attending staff for students during the medicine rotations.

(2) For a given time period there was exactly one form per student filled out by a house officer and one by an attending member of the academic staff.

(3) When there was more than one form completed for the same student, during the same time period, by the same evaluator type, one form was randomly selected for each evaluator type from that student.

The rationale for restricting the sample to one form per evaluator type for each student was to reduce the bias which may have been introduced if good students were evaluated more frequently than poor students. The decision to include only the forms from house officers and attending teachers in the internal medicine rotation was made to allow comparisons with the previous study (Dielman *et al.* 1980). The students' internal medicine rotation spanned more than one 4-week period, hence it was possible to receive more than one set of ratings during this rotation.

The sampling procedure resulted in a data set containing 1880 completed evaluation forms, 940 of which were completed by house officers and 940 of which were completed by attending teachers. These forms represented ratings of 303 individual students.

Data analysis

For each of the 12 periods, correlations were computed and the correlation matrices were factor analysed by the principle axes procedure. The number of factors was determined by the application of the Kaiser 'unity rule' (Guertin & Bailey 1970) and the Catell 'scree test' (Catell 1966). The 12 factor matrices were rotated by both orthogonal (Varimax) and oblique (Oblimin) procedures.

The internal consistency of the 'index

scores', computed by simple summation of the salient rating scales on each factor, and the total of the 13 scales were determined by computing Cronbach alpha coefficients. Separate Cronbach alphas were computed for the house officer and attending staff ratings across all 12 periods.

Interrater reliabilities for each of the 13 scales, the 'index scores' and total scores were computed separately for the house officers and attending staff. Interrater reliabilities were calculated by an ordinal data extension of the categorical data computation of the intraclass correlation coefficients (Landis & Koch 1977).

Results

The principle axes factor analyses resulted in two eigenvalues greater than 1·0 for four of the

THE UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
– Clinical Evaluation Form –

STUDENT NAME — LAST — FIRST — MIDDLE — OFFICE USE ONLY:

REPORT COVERS — FROM — TO — MO. DAY YR. — MO. DAY YR.

CLERKSHIP — ☐ REQUIRED CLERKSHIP — ☐ ELECTIVE CLERKSHIP — DEPARTMENT — SECTION — HOSPITAL

EVALUATOR'S NAME — LAST — FIRST INIT. — MIDDLE INIT.

☐1 ATTENDING STAFF ☐2 HOUSE OFFICER ☐3 PRECEPTOR ☐4 JOINT REVIEW ☐5 NURSE ☐6 SOCIAL WORKER ☐7 OTHER

INSTRUCTIONS: PLEASE SELECT THE MOST APPROPRIATE RATING FOR EACH CATEGORY. THE ENDS OF EACH SCALE LIST BEHAVIORS WHICH DESCRIBE POOR AND OUTSTANDING PERFORMANCE. A RATING AT EITHER END OF THE SCALE (1 OR 6) SHOULD BE ACCOMPANIED BY A WRITTEN COMMENT DOCUMENTING THE APPROPRIATENESS OF THE RATING. THE AVERAGE STUDENT'S PERFORMANCE WILL FALL WITHIN THE SHADED BOXES. MOST EVALUATIONS SHOULD BE WITHIN THESE BOXES.

HISTORY AND INTERVIEW							
1	☐0 NOT OBSERVED	☐1 INFORMATION IS INCOMPLETE OR INACCURATE; IMPORTANT INFORMATION IS MISSING	☐2	☐3	☐4	☐5	☐6 COMPREHENSIVE. INFORMATION IS THOROUGH AND PRECISE; DETAILED FOLLOW-UP OBTAINED
PHYSICAL EXAMINATION							
2	☐0 NOT OBSERVED	☐1 EXAMINATION INCOMPLETE; DEFICIENCIES IN TECHNICAL QUALITY; INACCURATE DATA	☐2	☐3	☐4	☐5	☐6 THOROUGH, TECHNICALLY SOUND EXAMINATION; OBTAINS ACCURATE DATA
DIFFERENTIAL DIAGNOSIS/PROBLEM LIST							
3	☐0 NOT OBSERVED	☐1 DIFFICULTY USING DATA; INADEQUATE PROBLEM LIST	☐2	☐3	☐4	☐5	☐6 ANALYZES AVAILABLE DATA; SYNTHESIZES INFORMATION; CONCISE, SUBSTANTIVE PROBLEM LIST
DIAGNOSTIC/THERAPEUTIC PLANNING							
4	☐0 NOT OBSERVED	☐1 PLAN IS INCOMPLETE OR INEFFICIENT; IMPORTANT TESTS ARE OVERLOOKED; DIFFICULTY INTERPRETING TEST RESULTS	☐2	☐3	☐4	☐5	☐6 PLAN IS COMPLETE AND EFFICIENT; INTERPRETS RESULTS CORRECTLY AND PRECISELY
PROCEDURAL SKILLS							
5	☐0 NOT OBSERVED	☐1 DIFFICULTY USING PROPER TECHNIQUES; DISORGANIZED; POOR COORDINATION	☐2	☐3	☐4	☐5	☐6 USES PROPER TECHNIQUES; ORGANIZED; PRECISE TIMING
KNOWLEDGE							
6	☐0 NOT OBSERVED	☐1 DIFFICULTY RECALLING AND RELATING BASIC SCIENCE AND CLINICAL INFORMATION	☐2	☐3	☐4	☐5	☐6 APPLIES BROAD BASE OF PERTINENT BASIC SCIENCE AND CLINICAL INFORMATION

Figure 1. The University of Michigan Medical School clinical evaluation form.

12 periods. In three of the remaining eight analyses, the second eigenvalue was greater than 0·94 and in all 12 analyses the scree test suggested the rotation of two factors. The first two eigenvalues, variances accounted for by the first two factors, and the correlations between the two factors resulting from the Oblimin rotations are shown for each of the 12 periods in Table 1.

The consistency of the factor patterns was examined by computing the congruence coefficients between factors (Harman 1960). The factor patterns were quite consistent as indicated by the matrix of congruence coefficients presented in Table 2. All but six of the 132 coefficients were 0·9 or greater, and all were 0·88 or greater.

SELF-EDUCATION						
7 ☐0 NOT OBSERVED	☐1 FAILS TO DEMONSTRATE KNOWLEDGE OF REQUIRED READING; DOES NOT ATTEND CONFERENCES, ROUNDS, ETC.	☐2	☐3	☐4	☐5	☐6 DEMONSTRATES KNOWLEDGE OF EXTENSIVE SUPPLEMENTAL READING; ATTENDS CONFERENCES, ROUNDS, ETC.
WRITTEN SKILLS						
8 ☐0 NOT OBSERVED	☐1 INCLUDES IRRELEVANT INFORMATION; MISSES IMPORTANT DATA; LATE	☐2	☐3	☐4	☐5	☐6 WRITE-UPS ARE PROMPT, CONCISE, THOROUGH AND ORGANIZED
ORAL PRESENTATIONS						
9 ☐0 NOT OBSERVED	☐1 DISORGANIZED AND POORLY INTEGRATED	☐2	☐3	☐4	☐5	☐6 COMPLETE, CONCISE, ORDERLY AND POLISHED
INTERPERSONAL RELATIONSHIPS WITH PHYSICIANS AND OTHER HEALTH PROFESSIONALS						
10 ☐0 NOT OBSERVED	☐1 DISRESPECTFUL AND UNCOOPERATIVE	☐2	☐3	☐4	☐5	☐6 RESPECTS AND COMPLEMENTS OTHER PROFESSIONALS; COOPERATIVE
INTERPERSONAL SKILLS WITH PATIENTS						
11 ☐0 NOT OBSERVED	☐1 LACKS COMMUNICATION SKILLS; CANNOT EXPLAIN THINGS; DOES NOT LISTEN TO PATIENTS	☐2	☐3	☐4	☐5	☐6 EMPATHETIC; RELATES WELL EVEN WITH DIFFICULT PATIENTS; LISTENS ATTENTIVELY
PROFESSIONAL RESPONSIBILITIES						
12 ☐0 NOT OBSERVED	☐1 NEEDS REPEATED REMINDERS OF ASSIGNMENTS; DOES LESS THAN PRESCRIBED WORK	☐2	☐3	☐4	☐5	☐6 PERFORMS DUTIES PROMPTLY AND EFFICIENTLY WITHOUT BEING REMINDED; WILLING TO SPEND ADDITIONAL TIME
OVERALL						
13 ☐0 NOT OBSERVED	☐1 INADEQUATE PERFORMANCE; FAIL	☐2	☐3	☐4	☐5	☐6 OUTSTANDING PERFORMANCE; HONORS
COMMENTS						
SIGNATURE:			**DATE:**			

FORM 3 - 7/61

Table 1. Eigenvalues, percentage of variance accounted for by two factors, *n*s, and factor correlations for 12 periods

Period	Eigenvalues I	Eigenvalues II	% variance	*n*	Factor correlations Oblimin rotation
1	8·78	1·18	76·6	174	−0·53
2	8·75	0·96	74·6	160	−0·65
3	9·52	0·86	79·8	182	−0·65
4	9·06	0·94	76·9	158	−0·62
5	9·37	0·83	78·5	182	−0·64
6	8·53	1·27	75·4	168	−0·53
7	9·50	0·79	79·1	150	−0·70
8	8·16	1·26	72·5	132	−0·46
9	8·94	0·98	76·3	152	−0·62
10	9·38	0·80	78·4	152	−0·70
11	9·10	0·70	75·4	144	−0·75
12	8·82	1·06	76·0	126	−0·51

The results of the 12 factor analyses are summarized in Table 3, which presents the median factor loadings and the range of loadings for each of the 13 variables across the 12 periods. These factor pattern values are those which resulted from the Oblimin rotations.

Factor 1 was given the label of 'problem-solving', consistent with the item content which included both procedural skills and cognitive abilities. The first factor received its highest loadings from ratings of abilities to acquire and utilize information to arrive at the appropriate diagnoses. These were items 1 (history and interview), 3 (differential diagnosis), 4 (diagnostic and therapeutic planning) and 6 (knowledge). Consistently high loadings were also contributed by items 2 (physical examination), 5 (procedural skills), 7 (self-education), 8 (written skills), 9 (oral presentations) and 13 (overall). While there were occasional departures, these items loaded more strongly on factor I than factor II in most of the 12 analyses. The remaining three items, 10–12 (interpersonal relationships with health professionals, interpersonal skills with patients, and professional responsibilities) loaded most highly on factor II in every case. The label attached to factor II was 'interpersonal skills'.

The internal consistency of the total of the 13 scales and the indices defining the two factors were examined by Cronbach alpha coefficients. The 'index scores' were computed by unit weighted summation of items 1–9 and 13 for the 'problem-solving' index, items 10–12 for

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	—	0·99	0·99	0·98	0·98	0·99	0·98	0·99	0·99	0·98	0·99	0·98
2	0·97	—	0·99	0·96	0·97	0·98	0·96	0·98	0·98	0·95	0·98	0·97
3	0·96	0·96	—	0·97	0·98	0·98	0·97	1·00	0·99	0·97	0·98	0·99
4	0·95	0·91	0·90	—	0·99	0·99	0·98	0·97	0·99	0·99	0·98	0·99
5	0·95	0·91	0·92	0·97	—	0·99	0·99	0·98	0·99	0·99	0·98	0·99
6	0·98	0·95	0·94	0·98	0·97	—	0·98	0·98	0·99	0·98	0·99	0·98
7	0·92	0·88	0·88	0·93	0·93	0·93	—	0·97	0·98	0·98	0·97	0·98
8	0·93	0·91	0·98	0·88	0·93	0·91	0·88	—	0·99	0·97	0·98	0·99
9	0·96	0·92	0·94	0·98	0·98	0·98	0·94	0·93	—	0·99	0·99	0·98
10	0·95	0·90	0·92	0·98	0·95	0·97	0·95	0·89	0·98	—	0·97	0·98
11	0·95	0·96	0·95	0·94	0·93	0·96	0·90	0·91	0·94	0·94	—	0·98
12	0·94	0·88	0·90	0·96	0·97	0·94	0·91	0·89	0·94	0·94	0·91	—

Note: Congruence coefficients for factor I appear above the diagonal; congruence coefficients for factor II appear below the diagonal (see Harman [1960], p. 257 for the computational formula).

Table 3. Summary of factor pattern values (medians and ranges) for the 13 items over the 12 periods

Item	Factor I Problem-solving	Factor II Interpersonal skills
(1) History and interview	0·82	0·06
	(0·51–0·91)	(−0·03–0·41)
(2) Physical examination	0·77	0·04
	(0·63–0·93)	(−0·03–0·26)
(3) Differing diagnosis/ problem list	0·87	−0·02
	(0·72–0·98)	(−0·10–0·20)
(4) Diagnostic/therapeutic planning	0·83	0·03
	(0·71–0·96)	(−0·09–0·21)
(5) Procedural skills	0·53	0·31
	(0·24–0·82)	(−0·10–0·51)
(6) Knowledge	0·83	0·05
	(0·70–0·90)	(−0·06–0·19)
(7) Self-education	0·70	0·18
	(0·62–0·84)	(0·01–0·29)
(8) Written skills	0·67	0·22
	(0·44–0·75)	(0·12–0·47)
(9) Oral presentations	0·70	0·19
	(0·41–0·75)	(0·10–0·43)
(10) Interpersonal relationships— professionals	0·11	0·81
	(−0·03–0·23)	(0·68–0·90)
(11) Interpersonal skills—patients	0·10	0·80
	(−0·03–0·19)	(0·70–0·88)
(12) Professional responsibilities	0·31	0·60
	(0·21–0·61)	(0·33–0·69)
(13) Overall	0·74	0·25
	(0·67–0·82)	(0·15–0·34)

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 may exceed unity (Hartman 1960, pp. 16–19).

the 'interpersonal skills' index, and all 13 items for the 'total clinical performance' index. The Cronbach alpha coefficients were computed separately for the ratings rendered by the house officers and by the attending teachers. These results are presented in Table 4. All Cronbach alphas were higher than 0·90, and did not differ substantially for house officers and attending teachers.

The interrater reliability coefficients for each item, the 'problem-solving' index, the 'interpersonal skills' index, and the 'total clinical performance' index are presented separately in Table 4. The 'problem-solving' and 'total clinical performance' scores were computed for those forms having non-missing ratings on at least 80% of the items included in the indices. Forms used to compute the 'interpersonal skills' score were required to have non-missing ratings on at least two of the three items included in the index.

The interrater reliability coefficients on the 13 items for the attending teachers ranged from 0·14 to 0·31, with a median coefficient of 0·24. The interrater reliability coefficients for the house officers on the 13 items ranged from 0·19 to 0·33, with a median coefficient of 0·24. The interrater reliability coefficients among house officers on the summed indices were 0·25 for the 'problem-solving' index, 0·28 for the 'interpersonal skills' index and 0·34 for the 'total clinical performance' index. The comparable interrater reliability coefficients for attending teachers were 0·11, 0·24 and 0·30.

Table 4. Cronbach alpha coefficients and interrater reliabilities for the two factors and the total scale as rated by house officers and attending teachers

	House officers (n=940)	Attending teachers (n=940)
Cronbach alphas		
Problem solving	0·96	0·97
Interpersonal skills	0·90	0·91
Total clinical performance	0·96	0·97
Interrater reliabilities		
(1) History and interview	0·24	0·24
(2) Physical examination	0·19	0·14
(3) Differential diagnosis/problem list	0·28	0·25
(4) Diagnostic/therapeutic planning	0·24	0·29
(5) Procedural skills	0·20	0·21
(6) Knowledge	0·27	0·24
(7) Self-education	0·25	0·27
(8) Written skills	0·27	0·24
(9) Oral presentations	0·24	0·31
(10) Interpersonal skills—professionals	0·22	0·15
(11) Interpersonal skills—patients	0·19	0·20
(12) Professional responsibilities	0·32	0·21
(13) Overall	0·33	0·29
Problem-solving	0·25	0·11
Interpersonal skills	0·28	0·24
Total clinical performance	0·34	0·30

Note: ns refer to the total number of ratings over the 12 periods.

Discussion

The consistency of the factor analytic results across the 12 periods suggests that the variables employed in the 13-item rating scale cluster reliably into two groups, which have been designated as the 'problem-solving' factor and the 'interpersonal skills' factor. These factors were replicable across the 12 factor analyses. This finding was consistent with the results of the analyses of the previous version of this form (Dielman *et al.* 1980). The internal consistency coefficients of the items comprising the two factors and the total scale were all 0·9 or greater, which is also similar to the findings in the earlier study, in which the alpha coefficients ranged from 0·83 to 0·95. As in the earlier study, there was very little difference in the internal consistency coefficients depending on whether the source of the ratings was house officers or attending teachers.

The interrater reliability coefficients for the items and indices were roughly comparable for the house officers and the attending teachers, which is contrary to the earlier results. In the 1980 study, the interrater reliability coefficients for the single items ranged from 0·22 to 0·37, with a median of 0·29 among attending teachers, while among house officers the coefficients ranged from 0·30 to 0·51, with a median of 0·36. In the current study, the single-item interrater reliability coefficients ranged from 0·19 to 0·33, with a median of 0·24 among house officers. Among attending teachers the range was from 0·14 to 0·31, also with a median of 0·24. In the earlier study the interrater reliabilities among house officers for the summed indices were 0·60, 0·44 and 0·61 respectively for the 'problem-solving', 'interpersonal skills' and total score, while the comparable interrater reliabilities among house officers in the current study were 0·25, 0·28 and 0·34. Among attending teachers, the interrater reliabilities for 'problem-solving', 'interpersonal skills' and the total score were 0·42, 0·36 and 0·40 in the earlier study. In the current study the comparable coefficients were 0·11, 0·24 and 0·30. The comparisons indicate that the inter-

rater reliabilities were somewhat lower in the current study compared to the 1980 study among both house officers and attending teachers, although the difference was greater among house officers.

As in the earlier study by Dielman *et al.* (1980), the results of the present study indicate that the dimensions of 'problem-solving' and 'interpersonal skills' as measured by the ratings of house officers and attending teachers, are highly replicable and internally consistent dimensions of medical students' clinical performance. The changes which were made in the scale, i.e. reduction in the number of items from 15 to 13 and the deletion of behavioural definitions from all scale points except the extremes, did not change these conclusions. The interrater reliabilities, especially among house officers, were lower in the current study. This may be a consequence of the deletion of the behavioural definitions from the intermediate scale points. The extent of the interrater agreement in the current study was 'fair' for both groups of raters, according to the admittedly arbitrary benchmarks employed by Landis & Koch (1977). This shortcoming, which is rather common in subjective judgements of human performance, can be compensated for by being aware of the problem and basing ratings to be used for summative purposes on the averages of judgements made by several observers, thus eliminating the error due to observer variation. This procedure, as shown by the results of Printen *et al.* (1973) can increase interrater reliability considerably. Another recommendation emerging from the comparison of the two studies is that performance rating scales include behavioural definitions which are as precise and as observable as possible, for as many points on the scale as possible, in the interest of increasing interrater agreement.

References

Cattell R.B. (1966) The scree test for the number of factors. *Multivariate Research* **1,** 245–76.

Davidge A.M., Davis W.K. & Hull A.L. (1980) A system for the evaluation of medical students' clinical competence. *Journal of Medical Education* **55,** 65–7.

Dielman T.E., Hull A.L. & Davis W.K. (1980) Psychometric properties of clinical performance ratings. *Evaluation and the Health Professions* **3,** 103–17.

Geertsma R.J. & Chapman J.E. (1967) The evaluation of medical students. *Journal of Medical Education* **42,** 938–48.

Gough H.G., Hall W.B. & Harris R.W. (1964) Evaluation of performance in medical training. *Journal of Medical Education* **39,** 679–92.

Guertin W.H. & Bailey J.P. (1970) *Introduction to Modern Factor Analysis.* Edwards Brothers, Ann Arbor, Michigan.

Harman H.H. (1960) *Modern Factor Analysis.* University of Chicago Press, Chicago.

Landis J.R. & Koch G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–74.

Printen K.J., Chappell W. & Whitney D.R. (1973) Clinical performance evaluation of junior medical students. *Journal of Medical Education* **48,** 343–8.

Received 20 *September* 1985; *accepted for publication* 28 *October* 1986