Factors affecting the reliability of an objective structured clinical examination (OSCE) test in neurology

R. F. GLEDHILL, D. CAPATOS

Summary

Inconsistencies in individual student scores from one year to the next prompted attempts to produce a more accurate objective structured clinical examination (OSCE) test in neurology.

A study of factors affecting the reliability of this test revealed that in spite of efforts to control patient and examiner variability, residual inaccuracies due to these effects remained. Also, the use of a uniform test structure may have led to student cueing. Student performance did not appear to be affected by the OSCE format *per se*.

Innovations that might improve test reliability in subsequent OSCEs were identified.

S Afr Med J 1985; **67:** 463-467.

In 1980, the Department of Internal Medicine at the University of Stellenbosch introduced an objective structured clinical examination (OSCE)[1] for in-course assessment of clinical skills.[2] Curriculum changes in 1981 resulted in the same student class being assessed in 1980, 1 month after the multisystem introductory course, and in 1981, 6 months after the clinical neurosciences course. For the neurology test, students were required to perform a physical examination (practical) and answer questions on their findings (written). Contrary to expectations, students recorded a significant average decline in practical (P) score (9,7%; $P < 0,0001$; $N = 141$) and written (Q) score (5,6%; $P = 0,015$; $N = 140$) between 1980 and 1981. These findings prompted a detailed investigation. Unless adequate explanations could be found and problems circumvented, doubts would have arisen about the advantages of the OSCE method[1] in comparison with traditional methods[3,4] of assessment.

On review, it was apparent that there were major differences in the manner in which the two OSCEs were designed and administered. As it happened, different individuals from different divisions of the Department of Internal Medicine devised the tests and examined on the two occasions. Also, in 1981 the time allowed for each part of the test was halved to 10 minutes and, in contrast to 1980, patient problems were diverse. Other differences that could have contributed to the decline in mean scores included the interval between training and assessment,[5-7] overall test complexity and criteria for rating student performance.

The most striking finding was the absence of significant correlation between the two examinations with regard to both individual P scores ($r = 0,12$) and Q scores ($r = 0,10$). Indeed, many students who did well in 1980 failed in 1981. We believe that this difference in the order of scoring within the student group is partly explained by patient and examiner variability, since the P score and Q score standard deviations in 1981 (19,9; 23,4) were larger than in 1980 (10,0; 17,0). An additional factor meriting consideration is that the order in which the students took the neurology test (test sequence) was dissimilar in the two examinations.

Accordingly, in 1982 more attention was paid to planning and control of the neurology test and marking consistency. Also, the test was designed and administered to facilitate a systematic study of factors affecting its reliability. The results of this study are presented to provide a further, objective appraisal of the OSCE method[8] and to report the formal evaluation of a test assessing clinical performance.[9]

Methods

1982 OSCE neurology test

The internal medicine OSCE format was similar to that described by Harden and Gleeson.[1] The examination comprised 20 stations and was completed over 5 consecutive days. Each day involved two morning sessions separated by a rest period of 45 minutes. At each session 14 - 16 students were examined, their grouping and sequence being determined alphabetically.

The content of the neurology test was derived from a consensus of faculty opinion on the level of competence expected from students in their 4th year (M.B. Ch.B. IV). Parallel patient problems were used. Students were required to examine lower limb motor function (practical) and to record their findings at the adjacent station (written). Five minutes were allowed at each station. The examiners were the four members of the neurology division who had devised the test. They were fully briefed on performance criteria to be assessed and relevant allocation of marks.[10]

For the practical part of the test, student performance was rated (P score) by each of a pair of examiners using a 10-item behavioural checklist.[11] Marks were awarded for thoroughness (8 items), overall proficiency (1 item) and attitude to the patient (1 item). A 3-point rating scale was used: 'optimum', 'satisfactory' and 'unsatisfactory' performance being awarded marks of 1, ½ and 0 respectively. On days 1 and 4, examiner W (R.F.G., principal specialist) was paired with examiner X (junior specialist), on days 2 and 5 with examiner Y (professor and departmental head), and on day 3 with examiner Z (senior medical officer). On day 5, examiner Y, using identical criteria, rated student performance by global judgement instead of using checklist marking. A different patient participated at

Department of Internal Medicine, University of Stellenbosch and Tygerberg Hospital, Parowvallei, CP
R. F. GLEDHILL, B.SC., M.B. B.S., M.R.C.P., M.D., *Senior Lecturer and Principal Specialist* (Present appointment: Head, Department of Neurology, Kalafong Hospital, Pretoria)
Institute for Biostatistics of the South African Medical Research Council, Parowvallei, CP
D. CAPATOS, M.MATH.

each session. All 10 patients were selected and evaluated by R.F.G. and judged to be of suitable disposition and to manifest physical signs that were reproducible and of comparable complexity.

For the written part of the test (Q score), there were 10 questions of the single true-false type. Only correct answers (1 mark) counted, but students were not advised whether counter-marking would be used. Identical questions, enquiring about the presence or absence of a physical sign, were used for all 10 patients. As their physical signs were not identical, the correct answers differed for the 10 patients. Answers were incorporated into the question paper which each student submitted on completion.

Statistical methods

Distribution-free pairwise tests (Mann-Whitney) were used in order to identify those examiners, sessions, etc. with mean scores significantly high or low in relation to the rest of the mean scores. No adjustments were made for multiple comparisons. Correlations of scores against time sequence were measured by Pearson's correlation coefficient (Spearman rank correlations gave substantially the same results). Estimates of test reliability were obtained by the split-half method, using the Spearman-Brown formula.[12] A standard significance level of 0,05 was used throughout.

Results

All but one of the 153 students in the M.B. Ch.B IV class participated in the internal medicine OSCE. Descriptive statistics of scores for each session of the neurology test are set out in Table I. For P score I (examiner W) there were no significant differences between mean scores at sessions A and B; such differences were present for P score II (examiners X, Y and Z) on day 3 only (examiner Z). By pairwise comparison of session averages, significant differences in mean scores were found at 3 sessions for P score I, at 9 sessions for P score II, at 6 sessions for P score$_{av}$ (average of P score I and P score II), and at 8 sessions for Q score. Mean P score I was significantly different from mean P score II in the first 6 sessions. Table I also shows the daily mean Q score which, apart from on day 3, increased serially and showed significant differences on all days except day 2.

There were significant differences between the overall mean marks awarded by examiners Y and Z and the other examiners (Table II). The correlation between P score and I and P score II was significant overall ($r = 0,51$; $0,01 < P \leqslant 0,05$) and at all sessions except 3A and 5A (Table III). There were 33 students to whom one examiner awarded a mark of 50% or less and the other a mark of over 50%. Details of these pass-fail judgements are illustrated in Fig. 1.

TABLE II. OVERALL MEAN OF MARKS AWARDED BY INDIVIDUAL EXAMINERS, WITH SIGNIFICANCE OF DIFFERENCES

Examiner	No. of students	Marks (%) Mean	SD
W	152	60,8	14,4
X	61	57,3	11,7
Y	60	66,1**(h)	15,6
Z	31	47,4**(l)	10,8
Average	152	57,9	12,8

**0,001 <$P \leqslant$ 0,01.
(h) = significantly high score; (l) = significantly low score.

A significant correlation between the order in each OSCE session in which students took the neurology test (test sequence) and test score was recorded for P score I in 3 sessions, for P score II in 2 sessions and for Q score in 1 session (Table IV). Estimates of reliability for the practical test overall were significantly different from zero for both P score I ($R = 0,61$; $P < 0,001$; $N = 152$) and P score II ($R = 0,53$; $P < 0,001$; $N = 121$). For the written test, estimates of reliability differed significantly from zero for the test overall and at 6 sessions; for 2 sessions the written test was altogether unreliable (Table V).

The overall mean P score$_{av}$ and mean Q score and the correlation between individual P score$_{av}$ and Q score are shown in Table VI.

Discussion and conclusions

Alphabetical grouping of students as used in the OSCE is not otherwise employed in the curriculum. In analysing the test

TABLE I. MARKS AWARDED BY EXAMINER W (P SCORE I) AND THE ALTERNATE EXAMINER (P SCORE II) FOR EACH SESSION AND WRITTEN (Q) SCORE FOR EACH SESSION AND DAY, WITH SIGNIFICANCE OF DIFFERENCES BETWEEN MEAN SCORES

		P score I	P score II		P score I minus		Q score (mean %)	
Session	No. of students	examiner W (mean %)	Examiner	Mean %	P score II (mean %)	P score$_{av}$ (mean %)	Session	Day
1A	15	59,3	X	51,7**(l)	7,6***	55,5*(l)	62,7*(l)	66,8*(l)
1B	16	58,1	X	50,6**(l)	7,5***	54,4*(l)	70,6	
2A	15	53,7**(l)	Y	66,7**(h)	−13,0***	60,2	62,7*(l)	69,3
2B	14	55,0**(l)	Y	65,4**(h)	−10,4***	60,2	76,4**(h)	
3A	15	64,7	Z	54,3*(l)	10,4***	59,5*(l)	59,3**(l)	66,5*(l)
3B	16	63,1	Z	46,9***(l)	16,2***	55,0*(l)	73,1*(h)	
4A	15	60,0	X	61,3*(h)	−1,3	60,7	75,3**(h)	75,0*(h)
4B	15	67,3**(h)	X	67,3**(h)	0,0	67,3**(h)	74,0**(h)	
5A	15	63,3	Y	64,7*(h)	−1,4	64,0*(h)	71,3	75,5*(h)
5B	16	63,1	Y	60,6	2,5	61,9	79,4**(h)	

* 0,01 < $P \leqslant$ 0,05.
** 0,001 < $P \leqslant$ 0,01.
*** $P \leqslant$ 0,001.
(h) = significantly high score; (l) = significantly low score.

TABLE III. CORRELATION BETWEEN P SCORE I AND P SCORE II FOR EACH SESSION, WITH SIGNIFICANCE OF THE CORRELATIONS

Session	No. of students	r
1A	15	0,75***
1B	16	0,71***
2A	15	0,76***
2B	14	0,75***
3A	15	0,34
3B	16	0,59*
4A	15	0,51*
4B	15	0,79***
5A	15	0,34
5B	16	0,65**

* $0,01 < P \leqslant 0,05$.
** $0,001 < P \leqslant 0,01$.
*** $P \leqslant 0,001$.
r = correlation coefficient.

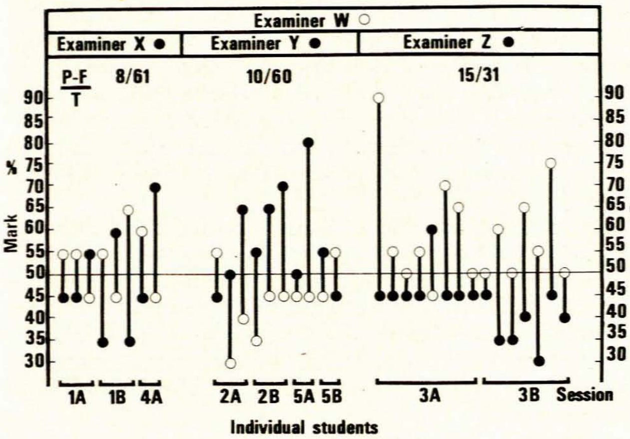

Fig. 1. Students ($N = 33$) for whom one examiner awarded a mark of ⩾ 50% and the other a mark of < 50%, with examiners and sessions ($\frac{\text{P-F}}{\text{T}}$ = pass-fail (P-F) as fraction of total (T) judgements of a pair of examiners).

TABLE IV. CORRELATION OF TEST SEQUENCE WITH P SCORE I, P SCORE II AND Q SCORE FOR EACH SESSION, WITH SIGNIFICANCE OF CORRELATIONS

		r		
Session	No. of students	P score I	P score II	Q score
1A	15	0,51*	0,01	0,61*
1B	16	0,10	0,21	0,00
2A	15	0,32	0,31	−0,24
2B	14	0,00	0,28	0,20
3A	15	0,32	−0,34	−0,24
3B	16	0,14	0,05	−0,39
4A	15	0,01	−0,53*	0,14
4B	15	0,54*	0,54*	0,06
5A	15	0,60*	0,03	0,12
5B	16	0,25	0,10	−0,42

*$0,01 < P \leqslant ,05$.
Test sequence = the order in the OSCE in which students took the neurology test; r = correlation coefficient.

TABLE V. ESTIMATES OF RELIABILITY (R) FOR THE WRITTEN PART OF THE TEST OVERALL AND AT EACH SESSION, WITH LEVELS OF SIGNIFICANCE

Session	No. of students	r	R
1A	15	0,46	0,53*
1B	16	0,67	0,80***
2A	15	0,49	0,66**
2B	14	0,13	0,23
3A	15	0,43	0,60**
3B	16	0,38	0,55*
4A	15	−0,36	I
4B	15	0,08	0,15
5A	15	−0,12	I
5B	16	0,64	0,78***
Overall	152	0,33	0,49***

* $0,01 < P \leqslant 0,05$.
** $0,001 < P \leqslant 0,01$.
*** $P \leqslant 0,001$.
r = correlation coefficient; $R = \frac{2r}{1+r}$ (see ref. 12); 1 = inconsistency in split-half test (negative r).

TABLE VI. OVERALL MEAN P SCORE AND MEAN Q SCORE AND CORRELATION BETWEEN INDIVIDUAL P SCORE AND INDIVIDUAL Q SCORE FOR 1980, 1981 and 1982

	No. of	P score (%)		Q score (%)		
Year	students	Mean	SD	Mean	SD	r
1980	144	64,5	10,0	65,1	17,0	0,25
1981	170	55,3	19,9	57,6	23,4	0,11
1982*	152	57,9	12,8	70,6	16,6	0,17

*P score$_{av}$.
r = correlation coefficient.

results, we have therefore assumed that no subgroup 'contamination effect' from previous class studies was operative and that student abilities were randomly distributed.

Test sequence effects

In the absence of patient and examiner variability, scores in a practical examination should provide an accurate account of student ability. The OSCE format itself, however, may affect the participants in such a way that scores for a particular test are not a true reflection of that student's ability. For example, some students experience greater pre-examination and intra-examination emotional tension than with other evaluation formats[13] and the repetitious demands may fatigue the student, patient or examiner.

Significant correlations between sequence and score were found in session 1A of the written part of the test and in sessions 1A, 4A (negative), 4B and 5A of the practical part. In seeking to explain the isolated finding for the written scores, it seems plausible that those students who presented early on in the initial session of the OSCE were disproportionately affected by anxiety, as has been noted in conventional oral examinations.[14] Had exposure to the OSCE format for the first time been a major factor, similar findings would have been expected in session 1B (from day 2, students taking the neurology test would know the OSCE format from examination in another specialty such as surgery or obstetrics). Any effect of test sequence on student performance can be judged only from scores in the written part of the test, since practical scores include the added factor of examiner variability. On the basis of the written scores we therefore conclude that, apart from the initial session, the order in the OSCE in which students took the neurology test had a negligible (trend) effect on their performance. In turn, we interpret this to imply that the OSCE format did not produce appreciable student fatigue.

The positive correlation between test sequence and P score I in sessions 1A, 4B and 5A suggests a tendency for examiner W to show bias towards the later students. This may reflect the effects of fatigue, since a similar bias was found in mean examiner ($N = 3$) grading in a loosely structured oral examination.[15] At this point, it should be mentioned that we do not believe that P score I and Q score in session 1A indicate the same phenomenon since, apart from the question of examiner variability and the fact that P score II in this session did not show this trend, there was no significant overall correlation between individual P score$_{av}$ and Q score. Although it is entirely feasible that examiner variability accounted for the practical test scores in session 4B, it is perhaps surprising that both examiners should have manifested such similar biases. An alternative explanation is that, by sheer chance, the ability of the first few students in session 4B was inferior to that of the last few. Finally, during session 4A, examiner X experienced frank symptoms of boredom and this was probably the factor most responsible for the significant negative correlation between test sequence and P score II in that session.

Patient effects

The most satisfactory method of eliminating patient variability in a practical examination involving 150 students is to use simulated patients.[16-18] As simulated patients were not available, we chose the next best alternative, namely patients with similar (i.e. parallel) problems.[19] In choosing to use one such patient for a single session only, we were influenced by the OSCE design, our obligations to the patients as willing participants, and our desire to minimize patient fatigue,[1,20] with its potential for intrapatient variability. (In so doing, we judged that 14 - 16 students was an adequate number to allow comparison of examiner rating). The choice of parallel problems generated the potential for intraclass cueing. To circumvent this problem by examining all students in one day[20] was not possible for administrative reasons. Although the factor of examiner variability precludes identification of any cueing effect in comparing day differences in practical scores, it is nevertheless of interest that the means of marks awarded by examiner W on day 5 were up to 5% higher than on day 1.

We consider that any effects produced by the variability of our patients would be manifested chiefly in the written scores, since student performance in the practical was rated for behaviours requiring only that patients be co-operative. The mean Q score differed significantly from at least one other in 8 sessions (the difference of about 20% between the lowest and highest subgroup score accords with Nowotny and Grove's[18] study of history-taking ability). Also, in 4 of the 10 sessions the estimate of reliability for the written part was not significantly different from zero. These findings suggest that, despite every effort to select patients with problems of equivalent complexity and with reproducible signs, test reliability was adversely affected by patient variability. One possible solution may be to select for future tests only those patients for whom high estimates of reliability in the written part have been recorded.

There was a serial increase in mean Q score on all but one of the 5 days. This suggests that the use of identical questions may have produced intraclass cueing. Since a similar trend was also noted in the practical scores, some doubt arises as to the benefits of using parallel problems and identical questions to aid test reliability. However, we think that it would be premature to discard their use before there is definitive evidence that any resulting cueing effects outweigh their value in providing comparable test situations for all students.

Examiner effects

Despite careful briefing as to performance criteria and allocation of marks, there were significant differences between the mean marks awarded by the pairs of examiners on days 1, 2 and 3. In session 3B this difference was as large as 16%, similar in magnitude to differences recorded for conventional clinical examination marking (15%)[21] but considerably in excess of differences for a comparable, structured examination (5,7%).[22] However, on days 4 and 5 differences between examiners were very much less (<2,5%). Discussion of student performance between examiners during the 5 days was deliberately avoided, but it is possible that non-verbal cueing or inadvertent remarks made by examiners W, X and Y — a 'contamination' effect[23] — could account for the smaller differences on the last 2 days. In addition, some form of 'practice' effect could have occurred, whereby judgements became less polarized with repetition. In this regard, it is noteworthy that 20 (60%) of the 33 pass-fail anomalies occurred in the first 5 sessions. Also, Wilson *et al.*[21] found that in the majority of cases marks did not differ by more than 5% on re-marking, compared with differences of up to 15% on first marking, and Colton and Peterson[15] recorded a tendency for increasing interexaminer reliability during a 3-day examination. If genuine, such 'contamination' and 'practice' effects could be put to future use by incorporating pilot sessions into pretest briefing.

The standard deviation of marks awarded by an examiner ranged from 10,8 to 15,6 over the four examiners, implying the need for more thorough pretest briefing. The extent of examiner variability is further revealed by comparisons between the overall mean of marks awarded by individual examiners, which show that a student rated by examiner Z would, on average, have received a score almost 19% lower than if rated by examiner Y.

It is interesting to note that the average mean mark awarded by examiner Y on day 2 (66,0%), using checklist marking, was little different from that on day 5 (62,6%), when using global judgement. Examiner Y was the senior and most experienced examiner and expressed a clear preference for global marking, citing the advantages of less distraction in judging a student's general proficiency and attitude to the patient and that the method was less fatiguing. However, correlation of marks awarded by examiner Y and examiner W was weaker for day 5 (0,34; 0,65) than for day 2 (0,76; 0,75), in keeping with the experience of others[24] when the two methods of marking are used.

On the 2nd day of participation, examiner Y experienced frank symptoms of boredom, with the need for a constant effort to maintain concentration. This is the most likely explanation for the weaker correlation between marks awarded by the two examiners in session 4A than in 4B. The experiences of examiners Y and X suggest that fatigue and boredom induced by the OSCE format could be an additional source of examiner variability. It is difficult to conceive how such effects could be contained further, given the limited demands imposed by the sectional design of the present OSCE, the need to avoid problems of validity associated with global judgement marking,[11] and the greater variability of marking when larger numbers of examiners are involved.[21]

The correlation between marks awarded by each of the pairs of examiners ranged from 0,34 at session 5A to 0,79 at session 4B. This range of values is much larger than that previously reported for structured clinical examinations, for either two[25] or more[11,19,22,24] examiners, and for conventional clinical examinations.[3,26] Furthermore, r values of 0,34 are much smaller than in many reported studies of both structured[19,22,24] and conventional[3,21,27] clinical examinations and are at variance with the degree of agreement in conventional oral examinations as assessed by estimates of concordance.[14] However, a similar range and degree of disagreement was found between three examiners in a loosely structured oral examination.[15]

At all sessions except 4B there were discrepancies of pass-fail judgements. Although such anomalies have been recorded in conventional clinical examinations[21,28] and a loosely structured oral examination,[15] the high proportion in this test (almost 22%) and the fact that 6 students awarded a distinction ($\geqslant$ 70%) by one examiner were failed ($<$ 50%) by the other are most disturbing. Although this aspect of interexaminer unreliability is clearly unacceptable, some consolation can be derived from the fact that almost half of these discrepancies derived from the lower marking of examiner Z compared with examiner W.

Whereas examiner reliability was not satisfactory — only on day 4 were the mean scores similar and the correlation high — that of the practical test itself was reasonably so, both when estimated from the marking of examiner W ($R = 0{,}61$; $P < 0{,}001$) and from the combined marking of the other three examiners ($R = 0{,}53$; $P < 0{,}001$). This contrast between estimated test and examiner reliability accords with the experience of Newble *et al.*,[8] who also found that the rater training did not resolve the outstanding problem of examiner reliability.[19] While in agreement with a similar study of a conventional examination,[28] these findings are in contrast to those recorded by Beswick *et al.*[22]

We consider that improvements in marking consistency in the neurology test may derive from more thorough preparation of examiners, although we agree with Newble *et al.*[19] that correct selection of examiners may be more important. Consensus rating could also provide a means of improvement, particulary when examiners include consistently 'high scorers' or 'low scorers',[28] as seems to have been the case with examiner Y (high) and examiner Z (low) in the present study.

OSCE neurology test, 1980-1982

Our cumulative experience with the OSCE neurology test is summarized in Table VI. The weak correlation between individual P score$_{\text{av}}$ and Q score in 1982 suggests that such results in 1980 and 1981 were probably recording the same phenomenon. The implication of these findings has been discussed elsewhere.[29]

It had been hoped that the use of parallel patients and fewer,[21] better briefed examiners would produce a more accurate test in 1982. Although mean score standard deviations were smaller in 1982 than in 1981, detailed analysis of the 1982 results revealed residual inaccuracies due to patient and examiner effects we had hoped to control. Nevertheless, we believe that the neurology test can be made more reliable if attention is given to the responsible factors identified in this study.

We wish to thank Professor J. Anderson, of King's College Hospital Medical School, London, for his criticisms of the manuscript, Professor H. P. Wassermann for his support, and Ms G. Muller for secretarial assistance.

REFERENCES

1. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979; **13:** 41-54.
2. Wassermann HP, Slabbert BR, Van Zyl JJW. Die objektief-struktureerde kliniese eksamen (OSKE). *S Afr Med J* 1982; **61:** 325-330.
3. Bull GM. Examinations. *J Med Educ* 1959; **34:** 1154-1158.
4. Stokes JF. *The Clinical Examination – Assessment of Clinical Skills* (Medical Education Booklet No. 2). Dundee: Association for the Study of Medical Education, 1974.
5. McGuire C, Hurley RE, Babott D, Butterworth JS. Auscultatory skill: gain and retention after intensive instruction. *J Med Educ* 1964; 39: 120-121.
6. Hazlett CB, Bachynski JE, Embleton J. Evaluation of on-campus continuing medical education programs in Alberta. *Can Med Assoc J* 1968; **98:** 674-676.
7. Bishop JM, Fleetwood-Walker P, Wishart E, Swire H, Wright AA, Green ID. Competence of medical students in history taking during the clinical course. *Med Educ* 1981; **15:** 368-372.
8. Newble DI, Hoare J, Elmslie RG. The validity and reliability of a new examination of the clinical competence of medical students. *Med Educ* 1981; **15:** 46-52.
9. Elstein AS, Lindenfield R. *A Compendium of Performance Evaluation Instruments for Health Professional Education: Report Submitted to the Division of Health Manpower Development.* Geneva: World Health Organization, 1979.
10. Harden RM, Stevenson M, Wilson Downie W, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975; **1:** 447-453.
11. Andrew BJ. The use of behavioural checklists to assess physical examination skills. *J Med Educ* 1977; **52:** 589-591.
12. Anastasi A. *Psychological Testing.* 4th ed. New York: Macmillan, 1976.
13. Kirby RL, Curry L. Introduction of an objective structured clinical examination (OSCE) to an undergraduate clinical skills programme. *Med Educ* 1982; **16:** 362-364.
14. Pokorny AD, Frazier SH. An evaluation of oral examinations. *J Med Educ* 1966; **41:** 28-40.
15. Colton T, Peterson OL. An assay of medical students' abilities by oral examination. *J Med Educ* 1967; **42:** 1005-1014.
16. Barrows HS. Simulated patients in medical teaching. *Can Med Assoc J* 1968; **98:** 674-676.
17. Newble DI, Elmslie RG, Baxter A. A problem-based, criterion-referenced examination of clinical competence. *J Med Educ* 1978; **53:** 72-76.
18. Nowotny RE, Grove DI. Description of an examination for the objective assessment of history-taking ability. *Med Educ* 1982; **16:** 259-263.
19. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ* 1980; **14:** 345-349.
20. Cushieri A, Gleeson FA, Harden RM, Wood RAB. A new approach to the final examination in surgery: use of the objective structured clinical examination. *Ann R Coll Surg Engl* 1979; **61:** 400-405.
21. Wilson GM, Lever R, Harden RM, Robertson JIS, MacRitchie J. Examination of clinical examiners. *Lancet* 1969; **i:** 37-40.
22. Beswick W, Cooper D, Whelan G. Videotape demonstration of physical examination: evaluation of its use in medical undergraduate teaching. *Med Educ* 1982; **16:** 197-201.
23. Stillman RM, Lane KM, Beeth S, Jaffe B. Evaluation of the student: improving the validity of the oral examination. *Surgery* 1983; **93:** 439-442.
24. Harper AC, Roy WB, Norman GR, Rand CA, Feightner JW. Difficulties in clinical skills evaluation. *Med Educ* 1983; **17:** 24-27.
25. O'Donohue WJ, Wergin JF. Evaluation of medical students during a clinical clerkship in internal medicine. *J Med Educ* 1978; **53:** 55-58.
26. Evans LR, Ingersoll RW, Smith EJ. The reliability, validity and taxonomic structure of the oral examination. *J Med Educ* 1966; **41:** 651-657.
27. Foster JT, Abrahamson S, Lass S, Girard R, Garris R. Analysis of oral examination used in speciality board certification. *J Med Educ* 1969; **44:** 951-954.
28. Ludbrook J, Marshall VR. Examiner training for clinical examinations. *Br J Med Educ* 1971; **5:**152-155.
29. Gledhill RF. Evaluating clinical competence in students. *Lancet* 1983; **i:** 595.