Medical Education, 1985, **19**, 238–247

Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence

J. J. NORCINI, D. B. SWANSON, L. J. GROSSO AND G. D. WEBSTER

American Board of Internal Medicine, University City Science Center, Philadelphia, Pennsylvania, U.S.A.

Summary. Despite a lack of face validity, there continues to be heavy reliance on objective paper-and-pencil measures of clinical competence. Among these measures, the most common item formats are patient management problems (PMPs) and three types of multiple choice questions (MCQs): one-best-answer (A-types); matching questions (M-types); and multiple true/false questions (X-types). The purpose of this study is to compare the reliability, validity and efficiency of these item formats with particular focus on whether MCQs and PMPs measure different aspects of clinical competence. Analyses revealed reliabilities of 0·72 or better for all item formats; the MCQ formats were most reliable. Similarly, efficiency analyses (reliability per unit of testing time) demonstrated the superiority of MCQs. Evidence for validity obtained through correlations of both programme directors' ratings and criterion group membership with item format scores also favoured MCQs. More important, however, is whether MCQs and PMPs measure the same or different aspects of clinical competence. Regression analyses of the scores on the validity measures (programme directors' ratings and criterion group membership) indicated that MCQs and PMPs seem to be measuring predominantly the same thing. MCQs contribute a small unique variance component over and above PMPs, while PMPs make the smallest unique contribution. As a whole, these results indicate that MCQs are more efficient, reliable and valid than PMPs.

Key words: *Clinical competence; *Education, medical, graduate; Psychometrics; Educational measurement/*methods; Internal medicine/educ; United States

Correspondence: John J. Norcini, PhD, American Board of Internal Medicine, University City Science Center, 3624 Market Street, Philadelphia, Pennsylvania 19104, U.S.A.

Introduction

The assessment of doctors' clinical competence has long been of concern to certifying and licensing agencies, educational institutions and the general public. Over the years, a number of clinical evaluation tools such as direct observation, chart audit, paper-and-pencil simulation, computer simulation, and patient simulation (see Barro, 1973 and Vu, 1979 for reviews of these approaches) have been developed. Due to factors such as cost in development and implementation, most of these approaches have proven impractical for large-scale use. Consequently, there remains heavy reliance on objective paper-and-pencil measures despite concern about their validity (Goran *et al.*, 1973; Feightner & Norman, 1976; Swanson *et al.*, 1982).

The most commonly used of the paper-

and-pencil methods are multiple choice questions (MCQs) and patient management problems (PMPs). Traditionally, it has been thought that MCQs measure knowledge, while PMPs, by virtue of the fact that they simulate a clinical encounter, measure knowledge plus problem-solving skills (Levine *et al.*, 1970; Levine, 1978). Studies investigating whether MCQs and PMPs measure different aspects of competence have produced variable results. In general, low to moderate correlations (0·2–0·5) between MCQs and PMPs have been observed (McGuire & Babbott, 1967; Schumacher *et al.*, 1974; Hubbard, 1978; Case, 1981), leading researchers to suggest that PMPs measure something different, although it is not clear what. Correlations between PMPs and external measures of clinical performance (usually ratings from clinical instructors) are also typically low to moderate, and approximately the same order of magnitude as correlations between MCQs and external measures (Schumacher *et al.*, 1974; Taylor *et al.*, 1976). Because these correlations are affected by the reliability of the measures used, it has been unclear if the low correlations imply that different traits are being assessed or that reliabilities (of the MCQs, PMPs and/or the external measures) are low. Findings of 'content/case/problem specificity' in the medical problem-solving literature (Barrows *et al.*, 1978; Elstein *et al.*, 1978; Swanson *et al.*, 1982; Norman *et al.*, 1983) and the low intercase correlations typically observed between PMPs (Donnelly *et al.*, 1974, 1982; Berner *et al.*, 1977; Mast *et al.*, 1982) both imply that the reliability of PMP-based scores will be quite low.

The study reported here examines the psychometric characteristics of PMPs and several MCQ formats, with particular focus on whether MCQs and PMPs are measuring the same or different aspects of clinical competence. More specifically, it compares the reliability, efficiency and simple and incremental validity of three MCQ formats (one-best-answer, matching and true/false questions) and PMPs.

Methods

Certifying examination structure and scoring

The data for the study are derived from the 1980, 1981 and 1982 American Board of Internal Medicine (ABIM) Certifying Examinations in General Internal Medicine, administered annually to 7000–8000 candidates. These examinations are fairly typical of licensure and certifying examinations. They consist of PMPs and three types of MCQs, one-best-answer questions (A-types), matching questions (M-types), and multiple true/false questions (X-types). The number of items of each type and the approximate testing time associated with each is displayed in Table 1 for the 3 years included in the study. Internal ABIM studies indicate that the average candidate completes thirty A- and M-type items and ninety-five X-type (i.e. nineteen multiple true/false questions with five items per question) items in an hour. Two and two-thirds of PMPs can be completed in an hour.

The PMPs are linear and moderately long (three to six problems and thirty to sixty options per case); they typically begin with a presentation of all relevant history and physical exam information and, in some cases, baseline laboratory values. Emphasis is on ordering laboratory tests and procedures and on medical management. Later problems within a case often introduce disease complications or a related disease. A-types include a mixture of short items measuring knowledge and more complex items with a lengthy 'case vignette' stem, requiring selection of a laboratory test, a diagnosis or a therapy. M-types usually involve matching of brief clinical stems (or diagnoses) with sets of laboratory findings. X-types are typically short, factual questions measuring knowledge of disease, laboratory test characteristics, drug complications, etc. All MCQ formats and PMPs frequently require interpretation of graphic material.

For each of the MCQ formats, 'number right' scores are calculated. Each of these

scores is then rescaled (standardized) using a linear transformation, so that the mean of a 'reference' group is 500 with a standard deviation of 100. This reference group consists of all candidates taking the certifying examination for the first time who graduated fom United States or Canadian medical schools with 3 years of approved residency training in general internal medicine and who completed the third year of graduate training within a year of examination administration. Non-reference group members are assigned scores by locating their raw scores in the reference group raw score distribution and assigning the associated standardized score. This approach is used to place all MCQ scores on the same scale.

Development of a composite PMP score is a multistep process. First, for each case, the test development committee assigns all options to one of six categories: (A) 'indicated and essential'; (B) 'indicated but not essential'; (C) 'neither indicated nor contraindicated'; (D) 'not indicated but not harmful'; (E) 'not indicated and expensive or somewhat dangerous'; and (F) 'not indicated and dangerous'. Second, each of the categories is assigned a numeric weight (typically, A=10, B=5, C=0, D=−2, E=−6, F=−12). Then, for each case and examinee, the numeric weights of the selected options are summed, yielding a raw score for each PMP. The raw scores are then individually standardized for each PMP to a mean of 500 and a standard deviation of 100, using the reference group procedure described above. The standardized PMP case scores are then summed and restandardized to a mean of 500 and a standard deviation of 100 for the reference group, yielding a composite PMP score on the same scale as the MCQ scores.

In order to form a composite examination score from the four standardized scores (A-types, M-types, X-types and PMPs) the item format scores are summed, with each score weighted in proportion to the time devoted to questions in that format. The resulting composite examination score is then restandardized so that, once more, the mean of the reference group is 500 and the standard deviation is 100. A composite MCQ score was also calculated in a fashion paralleling calculation of the examination composite score, but with the PMP contribution omitted.

Validity measures and subjects

As part of the certification process, evaluation forms are sent to the programme directors of all programmes in which a candidate participated during the years of residency training. The form requests that programme directors attest to the overall clinical competence of the candidate. Since 1980, these data have been collected on a 9-point rating scale (1–3 is the 'unsatisfactory' range—these candidates are not permitted to take the examination; 4–6 is the 'satisfactory' range; 7–9 is the 'superior' range). These ratings form the first validity measure; they represent an examination independent assessment of the clinical competence of candidates at the end of training. Subjects in this study are those taking the certifying examination for the first time in 1980, 1981 and 1982 who were rated on the 9-point scale.

In order to obtain a second validity measure of clinical competence reflecting the *quality of training* received, as well as relative performance in the training programme, 'high' and 'low' criterion groups were constructed. The 'high' criterion group consists of all first-takers who (1) graduated from United States or Canadian medical schools; (2) trained in selected high quality residency training programmes (these programmes were identified through peer nomination—five members of the ABIM, who were broadly familiar with internal medicine residency training programmes across the country, were asked to identify the best twenty-five programmes, and the thirty-five programmes receiving three or more votes were included in the high criterion group); and (3) were rated as 'superior' (7–9 on the rating form) by their programme director.

The 'low' criterion group consists of first-takers who (1) graduated from foreign medical schools; (2) did *not* attend one of the selected high quality residency programmes; and (3) were rated no higher than 'satisfactory' (4–6 on the rating form) by their programme director. These criterion groups are used as a dichotomous validity measure (0=low criterion group member; 1=high criterion group member); such extreme groups should provide a sensitive, powerful basis for investigation of validity.

Results

Reliability

Table 1 presents reliabilities for all scores in the years included in the study. For all MCQ formats, coefficient alpha was calculated, with M-types and X-types under a common stem treated as independent items. For the PMPs, coefficient alpha was also calculated, with standardized case scores treated as items. This approach to the calculations PMP reliability was used for two reasons. First, it parallels the domain definition used in test construction: cases are 'randomly sampled', while diagnostic and therapeutic selections within a case are not 'sampled' in any way. Second, this approach includes all available measurement information, yet avoids violating the assumption of item independence, since cases are independent. Reliabilities for the MCQ and examination composite scores were calculated using the Mosier formula (1943) for composite tests. The reliability of the examination composite is the same or slightly lower than that of the MCQ composite. Thus, the addition of several hours of PMPs does not improve overall reliability. This could be because MCQs and PMPs measure different and uncorrelated aspects of competence or that the PMP contribution to error of measurement balances its contribution to true score variance. Given the relatively low PMP reliability, the latter explanation seems more likely.

No reliability information is available for programme directors' ratings. Because the ratings are typically based on 3 years of contact between a candidate and programme director, with substantial input from other teachers, one would expect the ratings to be relatively reliable and valid, although there may well be differences in rater standards.

TABLE 1. Reliabilities and efficiences of item formats and composite scores

	A-types	M-types	X-types	MCQ composite	PMP composite	Exam composite
1980 examination						
Number of items	82	45	308	435	16	–
Testing time (hours)	2·7	1·5	3·2	7·4	6·0	13·4
Reliability	0·74	0·76	0·88	0·92	0·72	0·91
Efficiency	0·81	0·89	0·90	0·86	0·63	0·75
1981 examination						
Number of items	84	52	247	383	16	–
Testing time (hours)	2·8	1·7	2·6	7·1	6·0	13·1
Reliability	0·82	0·78	0·82	0·91	0·75	0·91
Efficiency	0·87	0·89	0·88	0·85	0·67	0·76
1982 examination						
Number of items	85	53	160	298	12	–
Testing time (hours)	2·8	1·8	1·7	6·3	4·5	10·8
Reliability	0·80	0·78	0·79	0·92	0·72	0·92
Efficiency	0·85	8·89	0·90	0·88	0·70	0·81

Efficiency

The reliability of the item format scores is heavily influenced by the number of items/testing time devoted to each. For planning test content and selecting item formats, efficiency—reliability per unit of testing time—is a more important measure of stability/homogeneity. Table 1 also presents efficiency coefficients—reliabilty coefficients corrected to 4 hours of testing time, the time allocated for an ABIM test booklet (the Spearman–Brown formula for the reliability of tests of different length; Lord & Novick, 1968). Of the item formats, A-, M- and X-types are the most efficient, while PMPs are least efficient. Because the examination composite is made up of these four scores, its efficiency is between the extremes.

Validity

Table 2 presents means and standard deviations for the item format and composite scores for the two criterion groups. Similar data for each level of programme directors' ratings are available on request. All scores rank order the groups in the expected direction, i.e. the high criterion group does better than the low criterion group, and groups with higher programme directors' ratings do better than those with lower ratings.

Taking the distance between groups as a rough indicator of score validity, it is apparent that the composite score is most discriminating, as one might expect, since it includes all measurement information available. Of the item formats, A-types, X-types and PMPs are about equally con-

TABLE 2. Means and standard deviations of item formats and composite scores by criterion group status

	Group size	A-types	M-types	X-types	MCQ composite	PMP composite	Exam composite
1980 examination							
Low criterion	551						
Mean		343	396	350	338	376	341
S.D.		148	134	135	144	138	146
High criterion	447						
Mean		539	533	547	547	540	548
S.D.		85	88	86	85	88	86
Combined criterion	998						
Mean		431	457	438	431	449	434
S.D.		157	134	151	160	144	160
1981 examination							
Low criterion	630						
Mean		357	398	344	343	342	326
S.D.		128	116	133	129	132	132
High criterion	515						
Mean		551	545	546	554	541	553
S.D.		80	86	82	79	83	76
Combined criterion	1145						
Mean		444	465	435	438	431	428
S.D.		146	127	151	152	150	158
1982 examination							
Low criterion	622						
Mean		319	367	337	315	309	294
S.D.		143	141	138	147	157	154
High criterion	486						
Mean		546	537	543	549	553	556
S.D.		82	87	86	82	76	78
Combined criterion	1108						
Mean		419	442	428	418	416	409
S.D.		165	147	156	169	176	181

TABLE 3. Correlations of item formats and composite scores with validity measures

	Group size	A-types	M-types	X-types	MCQ composite	PMP composite	Exam composite
1980 examination	4590						
Correlation with PDR†		0·31	0·28	0·33	0·34	0·30	0·35
Corrected correlation with PDR		0·32	0·30	0·33	0·33	0·28	0·32
	998						
Correlation with group status		0·62	0·51	0·65	0·65	0·57	0·64
Corrected correlation with group status		0·64	0·54	0·65	0·64	0·55	0·60
1981 examination	4907						
Correlation with PDR		0·35	0·29	0·32	0·35	0·33	0·37
Corrected correlation with PDR		0·36	0·31	0·33	0·34	0·31	0·34
	1145						
Correlation with group status		0·66	0·58	0·66	0·69	0·66	0·72
Corrected correlation with group status		0·67	0·61	0·67	0·68	0·64	0·68
1982 examination	4786						
Correlation with PDR		0·34	0·30	0·33	0·36	0·36	0·38
Corrected correlation with PDR		0·35	0·32	0·34	0·35	0·35	0·36
	1108						
Correlation with group status		0·68	0·57	0·65	0·65	0·69	0·72
Corrected correlation with group status		0·69	0·59	0·68	0·64	0·68	0·69

†PDR=Program directors' ratings

sistent in their ability to separate groups; M-types provide the least discrimination.

To verify these results numerically, Pearson product-moment corrrelations were calculated between programme directors' ratings and item format scores, along with point biserial correlations between item format scores and criterion group membership. The results of these analyses, shown in Table 3, document the trends noted in the pattern of means: composite correlations are consistently highest, while M-types are consistently lowest. A-types, X-types and PMPs produce roughly equivalent results, replicated across the 3 years. The absolute magnitude of the validity coefficients for prediction of programme directors' ratings is moderate (0·28–0·38) and the validity coefficients for the dichotomous criterion measure are large (0·51–0·72). These results provide strong positive evidence for the validity of the examination.

Since the magnitude of a validity coefficient is affected by score reliability, test length affects validity by way of reliability. Thus, Table 3 also presents the correlations between scores and validity measures which would be obtained if all scores were based on 4 hours of testing time, in other words, the validity per unit of testing time. Correlations with A-, M- and X-types all get larger since these scores were based on less than 4 hours of testing time. All other correlations get smaller. The efficiency of A- and X-types is clear in this analysis, with PMPs and M-types displaying poorer performance.

TABLE 4. Correlations among item formats and composite scores

	A-types	M-types	X-types	MCQ composite	PMP composite	Exam composite
1980 examination						
A-types	—	0·65	0·79	0·92	0·69	0·87
M-types	0·75	—	0·70	0·80	0·59	0·76
X-types	0·88	0·77	—	0·95	0·71	0·90
MCQ composite	0·95	0·85	0·97	—	0·75	0·95
PMP composite	0·79	0·69	0·82	0·84	—	0.92
Exam composite	0·92	0·81	0·94	0·97	0·95	—
1981 examination						
A-types	—	0·69	0·79	0·93	0·68	0·88
M-types	0·78	—	0·68	0·83	0·56	0·76
X-types	0·84	0·75	—	0·93	0·67	0·88
MCQ composite	0·95	0·87	0·95	—	0·71	0·94
PMP composite	0·76	0·68	0·74	0·79	—	0·91
Exam composite	0·91	0·83	0·90	0·96	0·94	—
1982 examination						
A-types	—	0·73	0·80	0·95	0·74	0·92
M-types	0·82	—	0·70	0·87	0·62	0·81
X-types	0·86	0·77	—	0·90	0·70	0·87
MCQ composite	0·97	0·91	0·93	—	0·76	0·96
PMP composite	0·82	0·71	0·80	0·84	—	0·92
Exam composite	0·95	0·86	0·91	0·97	0·94	—

Incremental validity

The previous analyses considered the relationship between validity measures and each score individually. A more important question is whether the scores are measuring the same or different aspects of clinical competence, since this presumption underlies the use of PMPs on many types of examinations. Test developers have asserted that although PMPs might be less reliable and efficient, they assess important aspects of competence not measured in the MCQ formats, thus justifying their inclusion because of *incremental* validity.

Table 4 presents intercorrelations among all scores for all examination years. Correlations above the diagonal are for the large group of first takers; correlations below the diagonal are for the combined high and low criterion groups (these correlations are higher because of the extreme groups). One of the original motivations for use of PMPs was a pattern of low to moderate correlations with MCQs, suggesting PMPs measure an aspect of clinical competence not assessed by MCQs. However, with these data, high correlations between MCQs and PMPs were observed in all examination years, indicating substantial overlap in what the two measure.

The magnitude of this overlap is underscored by correcting the correlations between the MCQ and PMP composites for attenuation (Lord & Novick, 1968). This correction predicts what the correlation would be if both measures were perfectly reliable. Correlations of 0·92, 0·86,and 0·93 are obtained for 1980, 1981 and 1982 respectively. The difference between the results of this study and others of PMPs and MCQs may be the relatively long PMP subtest on ABIM examinations. Such a subtest probably has a higher reliability than PMP scales used in other studies and thus the correlations were less attenuated.

Regression analysis was used to decompose the variation in validity measures into components accounted for uniquely by MCQs, uniquely by PMPs, and jointly by MCQs and PMPs. Table 5 presents the results of these incremental validity analy-

TABLE 5. Regression of MCQ and PMP composite scores on programme directors' ratings and criterion group status

	Proportion of explained variance	Breakdown of the explained variance		
		Shared by MCQs PMPs (%)	Unique to MCQs (%)	Unique to PMPs (%)
1980				
Programme directors' ratings	0·122	71	26	3
Group status	0·426	75	24	1
1981				
Programme directors' ratings	0·138	70	21	9
Group status	0·513	78	15	7
1982				
Programme directors' ratings	0·147	76	11	13
Group status	0·514	84	8	8

ses. The dependent variables are programme directors' ratings and criterion group membership; the independent measures are the composite PMP score and the composite MCQ score. As one might expect, the largest percentage of variance for all 3 years in both dependent measures is shared by both MCQs and PMPs and can be assessed by either one. MCQs contribute a small unique component over and above PMPs; in general, PMPs make a very small unique contribution. Regression coefficients for both MCQ and PMP predictors were positive in all analyses, but not of consistent magnitude, reflecting the high correlations between MCQs and PMPs.

Discussion

The purpose of this paper was to compare the psychometric characteristics of MCQs and PMPs. The results support the following conclusions:

(1) A-types, X-types and M-types are consistently more reliable and efficient than PMPs;

(2) A-types and X-types are more valid per unit of testing time than M-types and PMPs;

(3) MCQs and PMPs measure predominantly the same aspects of clinical competence; and

(4) MCQs make a small unique contribution to examination validity over and above PMPs.

These conclusions are, of course, tentative. It is possible that neither the programme directors' ratings nor criterion group membership are sufficiently sensitive indicators of clinical competence for good psychometric analysis, or perhaps PMPs measure some aspect of competence not reflected in these criteria. Further, the knowledge and skills examinees use in working through PMPs 'feel' more like the activities in which they engage when working with patients (McGuire & Babbott, 1967). It may be the case that with better measures of clinical competence, one might obtain different results.

The finding that PMPs did not perform well is consistent with the 'content specificity' of problem-solving skills (Barrows *et al.*, 1978; Elstein *et al.*, 1978; Norman, 1982; Swanson *et al.*, 1982). Doctors' performance varies considerably and non-systematically from one clinical situation to another. As a consequence, stable and generalizable scores can only result from extensive sampling of clinical situations. PMPs are at a considerable disadvantage since they are unable to sample as broadly as MCQs per unit of testing time.

It is conceivable that ABIM's PMPs are really 'glorified MCQs', and that more complex PMP-like formats (e.g. bran-

ching PMPs or uncued computerized PMPs) might measure aspects of clinical competence not assessed by MCQs. Such formats, however, are likely to exacerbate the content specificity problem, since they typically require more time per case than linear PMPs, and thus sample less broadly.

The results of this work, if confirmed through other investigations, point to an important implication for the use of paper-and-pencil measures of clinical competence. In less than a one-day testing situation, the inclusion of PMPs may produce less precise ability estimates. Further, since MCQs and PMPs measure predominantly the same aspects of clinical competence, future work might emphasize the evaluation of MCQ items designed to measure the application of knowledge in clinical situations.

Acknowledgement

This research was supported by, but does not necessarily reflect the policy or opinion of, the American Board of Internal Medicine.

References

Barro, A.R. (1973) Survey and evaluation of approaches to physician performance measurement. *Journal of Medical Education,* **48,** 1048–93.

Barrows, H.S., Feightner, J.W., Neufeld, V.R. & Norman, G.R. (1978) *Analysis of the clinical methods of medical students and physicians.* Report submitted to the Province of Ontario Department of Health and Physician's Services Inc. Foundation.

Berner, E.S., Bligh, T.J. & Guerin, R.O. (1977) An indication for a process dimension in medical problem-solving. *Medical Education,* **11,** 324–8.

Case, S.M. (1981) A new examination for the evaluation of diagnostic problem-solving. *Proceedings of the 20th Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Donnelly, M.B., Fleischer, D.S., Schwenker, J. & Chen, C.Y. (1982) Problem solving within a limited content area. *Proceedings of the 21st Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Donnelly, M.B., Gallagher, R.E., Hess, J.W. & Hogan, M.J. (1974) The dimensionality of measures derived from complex clinical simulations. *Proceedings of the 13th Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Elstein, A.S., Shulman, L.S. & Sprafka, S.A. (1978) *Medical Problem Solving: an Analysis of Clinical Reasoning.* Harvard University Press, Cambridge, Massachusetts.

Feightner, J.W. & Norman, G.R. (1976) Concurrent validity of patient management problems by comparison with the clinical encounter. *Proceedings of the 15th Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Goran, M.J., Williamson, J.W. & Gonnella, J.S. (1973) The validity of patient management problems. *Journal of Medical Education,* **48,** 171–7.

Hubbard, J.P. (1978) *Measuring Medical Education* (second edition). Lea & Febiger, Philadelphia, Pennsylvania.

Levine, H.G. (1978) Selecting evaluation instruments. In: *Evaluating Clinical Competence in the Health Professions* (eds M. Morgan & D. Irby). C. V. Mosby, St Louis, Missouri.

Levine, H.G., McGuire, C.H. & Nattress, L.W. (1970) The validity of multiple choice achievement tests as measures of competence in medicine. *American Educational Research Journal,* **7,** 69–82.

Lord, F.M. & Novick, M.R. (1968) *Statistical Theories of Mental Test Scores.* Addison-Wesley Publishing Co. Inc., Reading, Massachusetts.

McGuire, C.H. & Babbott, D. (1967) Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement,* **4,** 1–10.

Mast, T.A., Colliver, J.A., Anderson, M.B. & Solen, N.G. (1982) Validation of problem-solving measures: multitrait–multimethod matrix analysis. *Proceedings of the 21st Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Mosier, C.I. (1943) On the reliability of a weighted composite. *Psychometrika,* **8,** 161–8.

Norman, G.R. (1982) Medical problem-solving and the illusion of content specificity. *Professions Education Research Notes,* **4,** 10–1.

Norman, G.R., Feightner, J.W., Tugwell, P., Muzzin, L.J. & Guyatt, G. (1983) The generalizability of measures of clinical problem-solving. *Proceedings of the 22nd Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Schumacher, C.R., Burg, F.D. & Taylor, W.C. (1974) Computerization of a patient management problems examination to prevent retracing. *Proceedings of the 13th Annual Conference on Research in Medical Education.* Association of American Medical Colleges, Washington, D.C.

Swanson, D.B., Barrows, H.S., Friedman, C.P., Levine, H.G. & Norman, G.R. (1982) Issues in

assessment of clinical competence. *Professions Education Research Notes*, **4**, 2.

Taylor, W.C., Grace, M., Taylor, T.R., Fincham, S.M. & Skakun, E.N. (1976) The use of computerized patient management problems in a certifying examination. *Medical Education*, **10**, 179–82.

Vu, N.V. (1979) Medical problem-solving assessment: a review of methods and instruments. *Evaluation and the Health Professions*, **2**, 281–307.

Received 13 *June* 1984; *editorial comments to authors* 14 *September* 1984; *accepted for publication* 23 *October* 1984