A CRITERION-REFERENCED APPROACH TO MEASURING MEDICAL PROBLEM SOLVING

Validity of Patient Management Problems

FREDERIC M. WOLF
NANCY P. ALLEN
JAMES T. CASSIDY
BRUCE R. MAXIM
WAYNE K. DAVIS
University of Michigan Medical School

A criterion-referenced approach was used to examine the validity of a set of 15 Patient Management Problems (PMPs) representing a broad range of medical problems. Results of performance of 175 medical students indicated that a new problem-solving index that includes an open-ended differential diagnosis section had greater validity than the more traditional proficiency index, as the spread between the distributions of masters and nonmasters was greater for the problem-solving index. Implications for medical education and research are discussed.

AUTHORS' NOTE: An earlier version of portions of this article was presented at the 1983 annual conference on Research in Medical Education, Washington, DC. Address reprint requests to Frederic M. Wolf, Department of Postgraduate Medicine and Health Professions Education, University of Michigan, G1208 Towsley Center, Ann Arbor, MI 48109

EVALUATION & THE HEALTH PROFESSIONS, Vol. 8 No. 2, June 1985 223-240

Patient Management Problems (PMPs) have become an increasingly popular method of assessing clinical problem-solving competencies. PMPs are generally considered to possess adequate content validity, simulating the domain of knowledge, skills, and processes necessary to competently solve and manage patient cases (Hubbard, 1971; McGuire et al., 1976; Newble et al., 1981; Sedlacek and Nattress, 1972). As an evaluation procedure, they are preferred to mutiple-choice format examinations in approximating real-life medical problem-solving tasks. This is evidenced by their incorporation into Part III of the National Board Examinations (Hubbard, 1971). Additionally, their potential as an educational and training tool has been recognized (Feinstein et al., 1983; Marquis et al., 1984). However, the use of PMPs has given rise to some concerns, particularly in relation to their concurrent and construct validity (Goran et al., 1973; Newble et al., 1982; Norman and Feightner, 1981; Page and Fielding, 1980).

One of the most serious issues related to the concurrent validity of PMPs is a reported cueing effect on performance resulting from the listing of a menu of options from which to select (Goran et al., 1973; Newble et al., 1982; Norman and Feightner, 1981; Page and Fielding, 1980). Results of these studies are consistent with those of Goran et al. (1973) that "subjects performed significantly better on the PMP than in the clinical setting." However, several correlational studies do provide some evidence of the validity of PMPs. Hubbard (1971: 49) reported significant, positive correlations between PMP performance on Part III of the National Boards and Part II performance, ranging from .34 to .48. "These correlations, positive yet moderate, reflect the degree of correlation that would be expected between medical knowledge and additional elements of clinical competence inevitably based upon medical knowledge but representing skills that are to some degree independent of factual knowledge." Similarly, Newble et al. (1981) found significant relationships between performance on PMPs and final multiple-choice examinations in medicine (r_s= .54) and surgery (r_s= .62), supporting the validity of PMPs.

Several studies support the construct validity of PMPs by comparing the performance of subjects with varying degrees of

medical experience. Page and Fielding (1980) and Newble et al. (1981) found performance generallly to improve with increasing experience. Mazucca et al. (1981) provided evidence that resident's clinical judgment strategies on PMPs were more sophisticated and based on more selective data than were students'. Newble et al. (1982: 142), however, presented results that they believed did not support the construct validity of their PMP, asserting that "a test on which experienced doctors score less well than medical students cannot be considered a valid test of competence." However, a secondary analysis of their data revealed that physicians in their study did not perform significantly less well than medical students (Wolf, 1984). Unfortunately, the lack of power, small sample of subjects, and inclusion of only a single PMP in their study did not permit an adequate test of PMP construct validity.

Many PMP studies have been hampered by small samples of both subjects and PMPs. Although not directing their comments specifically to PMPs, Elstein et al. (1978: 86) have pointed out the varied and complex structure of medical problems: "(a) there is a high degree of content specificity in the process of solving medical problems; (b) it will therefore be difficult to predict from a physician's performance on one problem his performance on a new problem in a different domain; (c) content specificity implies that excellence in problem solving is content dependent as well as process dependent; and (d) assessing clinical competence is complex, and multiple measures will be required." Bashook (1976) noted that a single PMP or even several do not necessarily represent and assess the broader domain of medical problem solving. Bashook argued for a broad sampling of problem-solving processes, clinical disciplines, and content to evaluate clinical competence. Marshall (1977: 334) presented data to support this assertion, stating that there was not sufficient agreement on two PMPs "to allow only one problem to be used on the basis that it will measure sufficiently clearly a candidate's problem solving ability." Few, if any, of the previous studies of the validity of PMPs have attained this desired comprehensiveness.

The purpose of the present study was to examine the validity of patient management problems from a criterion-referenced perspec-

tive by sampling a wide range of medical problems with a large sample of medical students. Both Newble et al. (1978) and Feletti (1980) have pointed out the appropriateness of criterion-referenced (as opposed to norm-referenced) assessment of competence on PMPs, inasmuch as PMP items are clearly selected for their ability to determine whether a student has mastered a well-defined knowledge or skill domain. Norm-referenced tests are constructed specifically to maximize the variability of test scores and typically are used to rank order individuals on the measured ability. "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain" (Popham, 1975: 130). The terms "criterion-" and "domain-" referenced may be used interchangably according to Hambleton et al. (1978), although, for convenience, criterion-referenced has been preferred. Because criterion-referenced tests are not constructed to maximize test score variability, the distribution of scores on criterion-referenced tests tends to be more homogeneous. The criterion-referenced approach appears to be an ideal format for the type of knowledge measured by the PMP inasmuch as the domain of performance behaviors for the history, physical examination, diagnostic studies, differential diagnosis, and therapeutic procedures for a given standard medical case is relatively well defined. However, studies using norm-referenced approaches continue to dominate the literature, and few, if any, studies using a criterion-referenced approach have been reported.

Berk (1980) described several types of validity appropriate for criterion-referenced tests, two of which are addressed in the present study. Content validity refers to the content relevance or content representativeness of a test. Content validity is a concern during instrument development and is discussed further in the next section. Another type of validity of concern according to Berk is called decision validity. This type of validity addresses the use of test scores to classify individuals into either a "master" or "nonmaster" group. This is the type of validity addressed by the two hypotheses in the present study, which are described below. To the degree that differences between groups with different levels of skill and expertise such as masters and nonmasters can be

demonstrated, decision validity also may be thought of as evidence of construct validity.

By definition, performance by masters or skilled individuals on a valid criterion-referenced test should be significantly superior to performance by nonmasters on the test (Berk, 1980; Hambleton et al., 1978; Shepard, 1980). This statement leads to the first hypothesis tested in the present study: It is predicted that medical students will perform significantly better on PMPs after a problem-solving oriented educational intervention than before this intervention. This hypothesis should be supported if this curriculum intervention was effective in accomplishing its goals. A second research question compared the ability of two different PMP performance indexes to distinguish between masters and nonmasters. It is hypothesized that a new problem-solving index that included an open-ended differential diagnosis section should be more sensitive to differences between masters and nonmasters than the more traditional proficiency index, which does not include this type of section. The rationale for this hypothesis is that the proficiency index is based on selecting correct options or avoiding incorrect options from a menu of possible data gathering and therapeutic options. It has previously been demonstrated that this cueing effect leads to higher scores than scores obtained when PMPs are presented in an uncued, interview format (McCarthy, 1966; Newble et al., 1982). Although the problem-solving index in the present study also contains cued history, physical examination, diagnostic studies, and therapeutic procedures sections, the addition of an uncued differential diagnosis section is predicted to lead to greater discrimination between masters and nonmasters than the proficiency index, which includes only the cued sections.

METHOD

DESCRIPTION OF THE PMPs

Fifteen PMPs representing over 85 possible differential diagnoses were developed following a written linear format. Each

PMP began with an "opening scene" presenting the patient and the chief complaint. This was followed by 5 sections corresponding to the areas of: (A) history, (B) physical examination, (C) diagnostic studies, (D) the differential diagnosis and principal diagnosis, (E) therapeutic procedures. The first three sections (A, B, and C) each consisted of 15 to 40 decision options, accompanied by instructions to select the options that would be particularly important in establishing the diagnosis or in ruling out other plausible diagnoses. Descriptive feedback of the type normally generated clinically was provided for each option selected using a latent-image process. Section D contained open-ended questions asking for the diagnoses that were considered plausible in evaluating the patient and the most probable diagnosis. Section E resembled the first three sections of the PMP except that the latent-image process was not used to provide immediate feedback. This prevented students from working backwards and deducing the appropriate diagnosis.

The 15 PMPs used as instruments were based upon real patient cases selected to represent a broad range of content areas, as illustrated in Table 1. The 15 PMPs represent over 85 differential diagnoses related to eight organ system themes (cardiovascular, pulmonary, gastrointestinal, renal, neurologic, musculoskeletal, endocrine, reproductive), six medical specialties (medicine, surgery, pediatrics, neurology, psychiatry, and obstetrics-gynecology), and seven additional areas of medical practice (immunology, genetics, oncology, hematology, allergy, hypertension, infectious disease). Although this set of PMPs is not all-inclusive, it is fairly representative of the types of cases with which medical students could be expected to have some familiarity.

Worksheets following the format of the PMPs were constructed. Using these worksheets, the faculty member responsible for the medical content of each PMP worked with an educational specialist to develop a first draft. In developing the PMPs, faculty members began by delineating the differential and principal diagnoses for each case. Each of the information-gathering sections of the PMP was constructed to allow the informed student to systematically eliminate competing diagnoses and focus more confidently on the most probable diagnosis.

TABLE 1
Summary of Differential Diagnoses for 15 Patient Management Problems (PMPs)

PMP	Differential Diagnosis
1	Allergic asthma, Allergic reaction, Allergic rhinitis, Aspirgillosis, Emphysema, Triad asthma
2	Acute cholecystitis, Acute pancreatitis, Appendicitis, Food poisoning, Gastroenteritis, Peptic ulcer, Renal calculus
3	Diabetes insipitus - central, Diabetes insipitus - nephrogenic, Diabetes insipitus - psychogenic, Diabetes mellitus, Urinary tract infection
4	Ankylosing spondylitis, Psoriasis, Rheumatoid arthritis, Systemic lupus erythematosus
5	Bacterial lung abscess, Cancer, Crypotococcus infection, Dermatomyositis, Fungal infection, Mycobacterial infection, Pulmonary infarct
6	Coarctation of the aorta, Essential hypertension, Pheochromocytoma, Primary aldosteronism, Renal artery stenosis, Renal hypertension
7	Alcoholic cirrhosis with acute decompensation, Benign stricture of distal common bile duct, Choledocholithiasis, Common bile duct obstruction, Drug related jaundice, Hepatitis, Pancreatic cancer
8	Infectious arthritis, Rheumatoid arthritis, Seizure disorder, Systemic lupus erythematosus
9	Alzheimer's disease, B12 deficinecy, Bilateral subdural hematomas, Creutzfeldt - Jakob disease, Depression, Drug intoxication, Huntington's disease (chorea), Korsakoff's psychosis, Normal pressure hydrocephalus, Tertiary syphillis
10	Heart block, Dissecting thoracic aneurysm, Hiatus hernia with reflux, Myocardial infarction, Pericarditis, Pneumothorax, Pulmonary embolus
11	Congenital adrenal hyperplasia, Hermaphrodite, Hydroxylase deficiency, Maternal androgen/progesterone excess, Partial testicular feminization, Peno-scrotal hypospadius, 5 α reductionase deficiency
12	Herpes simplex encephalitis, Left middle cerebral artery stroke, Left parietal lobe tumor, Left parietal lobe abscess, Left-sided chronic subdural hematoma
13	Diabetic nephropathy, Glomerulonephritis, Myeloma, Primary hyperparathyroidism
14	Malpositioned ET tube, Patent foramen ovale, Pneumothorax
15	Amenorrhea, Androgens axcess, Anorexia nervosa, Hypothalamic amenorrhea, Pituitary adenoma, Polycystic ovarian disease Pregnancy, Thyroid dysfunction

Each PMP was edited and revised by an educational specialist-nurse clinician team. One purpose of this revision was to ensure that each section contained a balance of positively- and negatively-scored options and that enough options were included to distract the uninformed student. Furthermore, distractor options had to allow the uninformed student to act on the same set of misconceptions throughout the PMP. A second purpose of the revision process was to ensure that the options in each of the information-gathering sections of the PMP were presented uniformly across PMPs. For instance, the options in the history section of the PMPs were consistent with categories of information contained in the history database portion of the health status examination as presented in the Medical School's curriculum (Wolf et al., 1982). This uniformity minimized the potential cueing effect of the written PMP and, in addition, increased the effectiveness of the PMP for educational purposes. The final version of each PMP was reviewed by the clinical faculty member responsible for its content to ensure that the PMP was accurate and at the level of a second-year medical student. Additional description of the development of these PMPs is presented elsewhere (Allen et al., 1982).

SCORING OF THE PMPs

Each decision option in sections A, B, C, and E was scored positive (appropriate or indicated), negative (inappropriate or contraindicated), or zero (neutral). One point was given for each positive option selected and each negative option avoided. Responses to neutral options did not influence an individual's score. Each of the appropriate differential diagnoses included in Section D received one point. An additional point was awarded for identifying the most probable diagnosis from among the differential diagnoses. Although different weighting systems have been used in scoring PMPs, the present system was selected because of its relative simplicity and because studies comparing unit weighting to other commonly used methods have found the scores to be relatively comparable. Bligh (1980a, 1980b, 1981) found the scores to be "similar enough to be used interchangeably.

No preference for any one system is suggested by the statistical indices of score quality that were examined." Correlations among different scoring systems in the Bligh studies were in the .80s and .90s, suggesting there was not much difference in the rank order of medical students as a function of the different scoring systems. Donnelly (1976: 162) compared the +1, 0, –1 weighting method to a method that weighted items from +8 to –8. He found little difference between the two methods, although he noted that the "the more simple weighting (–1, 0, +1) method appears to be more reliable." In addition, Donnelly (1976: 164) suggested that with the more simple weighting method, scores are "more easily interpretable" and "the physician's rating task is simplified." Given no apparent superiority of one system over another, parsimony dictated that the simplest scoring system be employed in the present study.

In scoring the PMPs, each section of the PMP was weighted equally. A problem-solving index was used for grading student performance. The problem-solving index consisted of the average of the percentage of correct points across the five sections of each PMP. A proficiency index for each PMP was calculated by summing the total number of positive options correctly selected and negative options correctly avoided and dividing by the total number of possible correct points. These indices are identical with the exception that the problem-solving index includes the open-ended differential diagnosis section, whereas the proficiency index does not. This scoring system for the cued sections is similar to that used by the National Board of Medical Examiners Part III PMPs (Hubbard, 1971), as well as that used by the American Board of Internal Medicine certifying examination (Bollet, 1981). Both the problem-solving and proficiency indices were used to examine the research question regarding their validity to distinguish between masters and nonmasters. These scoring algorithms differ slightly from those provided by McGuire et al. (1976) in that the denominator in their algorithms consists of only the total number of possible appropriate (positive) choices, whereas the denominator in the present algorithms also includes the total number of possible inappropriate (negative) selections avoided. It is equally important to avoid inappropriate procedures as to

perform appropriate procedures when caring for patients. Unnecessary intrusion in patients' lives, added risks, and possible side-effects, as well as added expenses and nonoptimal use of resources may result from unnecessary or contraindicated procedures.

Cronbach alpha internal consistency reliability coefficients were .515 (4 pretest scores) and .742 (11 posttest scores) for the problem-solving index and .522 (4 pretest scores) and .737 (11 posttest scores) for the proficiency index. It is apparent from examining the magnitude of these coefficients that nonmaster performance is less reliable or consistent than master performance, an intuitively logical finding. Nonmasters would be unlikely to exhibit very consistent behavior on measures for which they are not trained or skilled. Masters, on the other hand, could be expected to perform more consistently. The fact that these reliability coeffficients were based on only 4 PMPs for the pretest in contrast to 11 PMPs for the posttest also likely contributed to the greater reliability of the posttest. These posttest reliability coefficients (for masters) compare favorably with those reported by Bligh (1980b) for a sample of senior medical students, and are adequate to warrant making inferences regarding group performance in the present study.

SUBJECTS AND PROCEDURE

The fifteen patient management problems were administered to 175 medical students as a required part of an experimental four-week "Interphase Program." The Interphase Program occurred at the close of the second year and was designed to provide a transition between the first two preclinical years of the Medical School curriculum and the clinical clerkships. Its expressed purpose was to introduce students to the concepts of clinical problem solving, ward experience, and the integration of concepts with facts. Students were introduced to problem solving and the integration of basic science knowledge with clinical concepts in a uniform sequence containing clinical problem-solving conferences, longitudinal clinical problems, pathophysiologic correlations of clinical problems, and seminars discussing state-of-the-art topics that were developed around eight organ-system

themes. An additional sequence consisted of ward, clinical skills, and psychiatric examination rotations. The major objective of the use of PMPs in this new program was the evaluation of students' skills in clinical problem solving. Each of the four weeks of Interphase focused on two of eight organ-system themes: Cardiovascular-Pulmonary (Week 1), Gastrointestinal-Renal (Week 2), Neurologic-Musculoskeletal (Week 3), and Endocrine-Reproductive (Week 4). Four PMPs were administered as a 2-hour pretest two weeks prior to Interphase. Each PMP in the pretest corresponded to one of the four weeks of Interphase and was randomly selected from among the PMPs for that week. Two PMPs corresponding to the two organ-system themes of the week were given as a posttest at the end of each week of Interphase. At the conclusion of the fourth week of Interphase, in addition to the two Week 4 posttest PMPs, students took three additional PMPs corresponding to organ-system themes from the first three weeks of Interphase. Additional description of the Interphase Program is presented elsewhere (Allen et al., 1984).

DESIGN AND ANALYSES

Students' average performance across all four pretests and all eleven posttests served as the PMP measures. This averaging procedure was employed based on the assertion that a wide range of PMPs—rather than individual PMPs—is necessary to assess the broader construct of medical problem solving (Bashook, 1976; Elstein et al., 1978; Marshall, 1977). A criterion-referenced approach was taken to assess the validity of the PMPs. According to Hambleton, et al. (1978: 15), there should be performance differences between " 'masters' of the material included in a test (perhaps a group of examinees after instruction) and those who would be expected to be 'non-masters' perhaps a group of examinees prior to receiving instruction." Therefore, students' pretest-posttest performance was compared using paired t-tests and the Wilcoxan Matched-Pairs test for change in performance (Conover, 1980) on the problem-solving and proficiency indices.

Measures of effect size (Cohen, 1977) were used to determine the magnitude of changes in performance to accompany measures

of their direction and statistical significance provided by the formal statistical tests. Cohen's d was used to provide this effect size index. According to Cohen, effect sizes may be interpreted as small (d = .2), medium (d = .5), and large (d = .8). Additionally, two measures of nonoverlap between the pretest and posttest distributions based on d were used (Cohen, 1977). U_1 represents the degree to which the two distributions overlap or are superimposed upon each other, and U_3 represents the percentage of the pretest distribution that the upper half of the posttest distribution exceeds. Assuming that student performance on the pretest PMPs before the Interphase curriculum represents nonmaster performance, and posttest PMP performance represents master performance, a valid criterion-referenced test should exhibit as little overlap as possible between the distributions of masters and nonmasters (Berk, 1980; Hambleton et al., 1978; Shepard, 1980).

RESULTS

The pretest and posttest distributions were approximately normal as indicated by measures of skewness and kurtosis. The assumption of homogeneity of variance also was satisfied as there were no significant differences between variances. Therefore, the assumptions underlying the t-tests and effect sizes were tenable for these data. Changes in performance from pre- to posttesting are summarized in Table 2.

Improved PMP problem-solving scores were evidenced by 171 of 175 students (Wilcoxan $z = 11.44$; $p < .001$) on the problem-solving index. Mean scores on the problem-solving index improved significantly ($p < .001$) from an average of 59% correct on the pretest to an average of almost 68% correct on the posttest. This translated into an improvement of 2.30 standard deviation units (d). Put slightly differently, the average student's performance on the posttest was equivalent to performance at the 98.9th percentile of the pretest distribution (U_3) on problem solving. Approximately 86% of the two distributions did not overlap (U_1). The fact that only 14% of these distributions did overlap supports the construct (decision) validity of the PMPs based on the

TABLE 2
Results of Paired t-Tests and Effect Sizes for Average Pre- and Posttest PMP Performance for 175 Medical Students

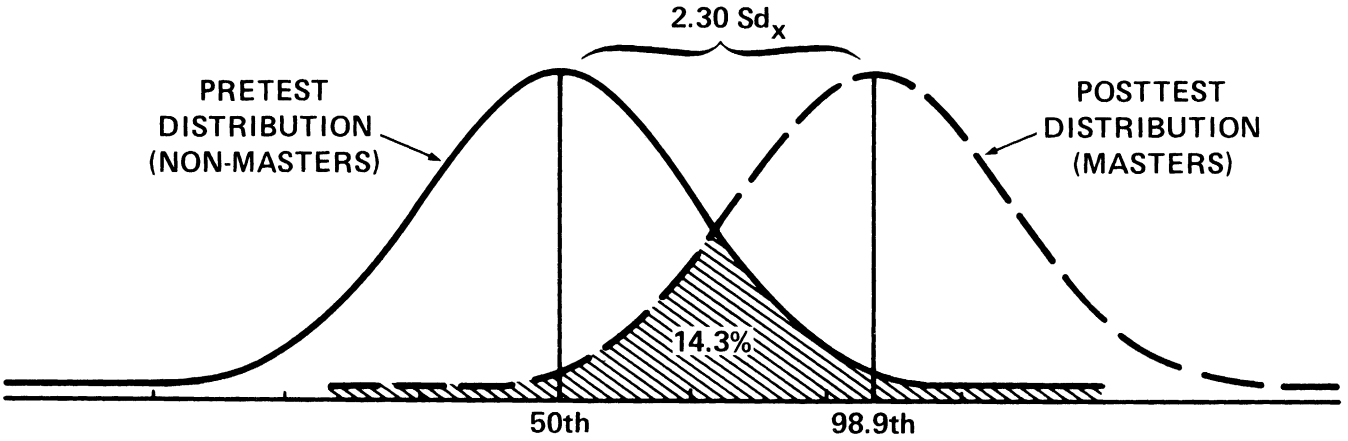
PMP Index	Pretest Mean	Posttest Mean	Within Sd	t	p<	d	U_1(%)	U_3(%)
Problem-Solving	59.02	67.86	3.85	-30.37	.001	2.30	85.7	98.9
Proficiency	71.24	73.24	4.05	- 6.52	.001	0.49	32.4	68.7

nonoverlap between masters and nonmasters. These findings are summarized visually in Figure 1.

In contrast to the findings for the problem-solving index, 121 students improved and 54 declined in performance on the proficiency index. Average proficiency performance improved significantly ($t = 6.52$; $p < .001$) from slightly over 71% correct on the pretest to slightly over 73% correct on the posttest. This was approximately a half standard deviation improvement ($d = .49$). The average proficiency performance on the posttest was equivalent to performance at the 68.7th percentile of the pretest distribution (U_3). Although 32% of the two distributions did not overlap (U_1), there was 68% overlap between the distributions. Comparing these findings in Figure 2 with those in Figure 1 for the problem-solving index, it can be seen that the problem-solving index evidences greater spread between masters and nonmasters and therefore greater construct (decision) validity than the proficiency index.

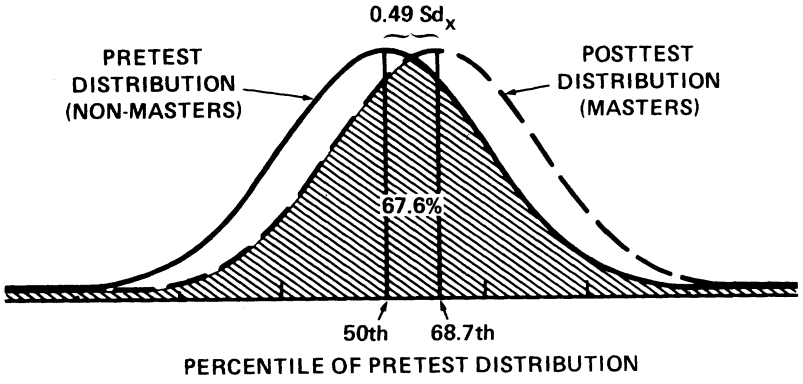
DISCUSSION AND IMPLICATIONS

Significant improvement on the PMP problem-solving and proficiency indices following the problem-solving oriented curriculum supported the construct (decision) validity of these PMPs. Results suggested that the problem-solving index has greater validity than the proficiency index as the spread between the distributions of masters and nonmasters was greater for the



NOTE: Average effect size in standard deviation units (Sd_x) and measures of nonoverlap between master and nonmaster distribution.

Figure 1: Medical Student Performance on the PMP Problem-Solving Index



NOTE: Average effect size in standard deviation units (Sd_x) and measures of non-overlap between master and nonmaster distributions.

Figure 2: Medical Student Performance on the PMP Proficiency Index

problem-solving index. PMPs often are developed to assess clinical problem solving because they provide a higher fidelity simulation of clinical skills than other examination formats such as multiple choice tests. Therefore, it seems logical for PMPs to include a differential diagnosis section. Evidence of the validity of a problem-solving index that includes this type of section has been presented. This index appears particularly sensitive to growth in students' diagnostic reasoning abilities. The more commonly used proficiency index appears to be less sensitive to this change because it does not include an uncued differential diagnosis section. These findings are strengthened by the large sampling of both subjects and medical problems, a deficiency in many previous studies. Although the pretest-posttest design used in this study provides initial evidence of the validity of these PMPs for assessing medical problem solving, an experimental-control group design would be even stronger. Studies employing this more rigorous design are recommended and at least one is currently in progress.

It should be noted that this study represents an initial examination of the validity of these particular PMPs. According

to Cronbach and Meehl (1955: 291), "to validate a claim that a test measures a construct, a nomological net surrounding the concept must exist." This nomological network consists of specific testable hypotheses that should lend support to the confirmation or disconfirmation of the validity of the construct. The hypothesis tested and supported in the present study is that a valid criterion-referenced PMP performance index should yield higher scores for masters than for nonmasters. In this regard results for the problem-solving index were superior than for the proficiency index. Other testable hypotheses pertaining to the construct and criterion-related (concurrent and predictive) validity of these PMPS need to be advanced and empirically examined. One such hypothesis is that PMP problem-solving performance is related to actual behavior in clinical problem-solving with patients. Examination of this and other testable hypotheses should help to clarify further the issue of the validity of PMPs. One of the few studies of this nature was reported by Donnelly and Prevot (1978). They examined the degree to which choices made by family medicine residents on PMPs reflected the differential effects found between emergency and nonemergency type patient cases, such as "physicians in real-life emergency situations asked fewer, more specific, and more useful questions than in the real-life non-emergency situation." They interpreted resident performance on five emergency and five nonemergency PMPs "as providing evidence for the construct validity of the PMPs" (Donnelly and Prevot: 60). A more recent study provided evidence that medical residents who ordered more services for their real clinic patients also tended to do so on PMP simulations (White et al., 1984). Additional studies of this nature clearly are desirable.

REFERENCES

ALLEN, N. P., J. T. CASSIDY, J. M. GARRISON, F. M. WOLF, and W. K. DAVIS (1984) "The design and evaluation of an interdisciplinary "Interphase" curriculum for second-year medical students." Submitted for publication.

ALLEN, N. P., J. M. GARRISON, J. T. CASSIDY, F. M. WOLF, W. K. DAVIS, and J. G. CALHOUN (1982) The Development of Patient Management Problems to Assess

Clinical Problem-Solving Knowledge and Abilities of Second-Year Medical Students. [Abstract] Innovations in Medical Education Exhibits. Washington, DC: Association of American Medical Colleges.

BASHOOK, P. G. (1976) "A conceptual framework for measuring clinical problem-solving." J. Medical Education 51: 109-114.

BERK, R. A. (1980) Criterion-Referenced Measurement: The State of the Art. Baltimore: Johns Hopkins Univ. Press.

BLIGH, T. J. (1981) Written simulation scoring: An in-depth look at three common systems. Abstracts of Papers and Symposia of the Annual Meeting of the American Educational Research Association, Los Angeles.

———(1980a) "The effects of various weighting schemes in the scoring of patient management problems in medicine." Ph.D. dissertation, University of Illinois, Urbana.

———(1980b) "Written simulation scoring: a comparison of nine systems." Abstracts of Papers and Symposia of the Annual Meeting of the American Educational Research Association, Boston: American Educational Research Association.

BOLLET, A. J. [Ed.] (1981) Harrison's Principles of Internal Medicine Patient Management Problems: Pretest Self-Assessment and Review. New York: McGraw-Hill.

COHEN, J. (1977) Statistical Power Analysis for the Behavioral Sciences (rev. ed.). New York: Academic.

CONOVER, W. J. (1980) Practical Nonparametric Statistics (2nd ed.) New York: John Wiley.

CRONBACH, L. J. and P. E. MEEHL (1955) "Construct validity in psychological tests." Psych. Bull. 52: 281-302.

ELSTEIN, A. S., L. S. SHULMAN, S. A. SPRAFKA et al. (1978) Medical Problem Solving. Cambridge, MA: Harvard Univ. Press.

DONNELLY, M. B. (1976) "Measuring performance on patient management problems," pp. 161-166 in Proceedings of the Fifteenth Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges.

———and E. L. PREVOT (1978) "Construct validity of patient management problems: Emergency versus non-emergency contexts," pp. 57-62 in Proceedings of the Seventeenth Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges.

FELETTI, G. I. (1980) Reliability and validity studies on modified essay questions. J. Medical Education 55: 933-941.

FEINSTEIN, E., L. P. GUSTAVSON, and H. G. LEVINE (1983) "Measuring the instructional validity of clinical simulation problems." Evaluation and the Health Professions 6: 61-76.

GORAN, M. J., J. W. WILLIAMSON, and J. S. GONNELLA (1973) "The validity of PMPs." J. Medical Education 48: 171-177.

HAMBLETON, R. K., H. SWAMINATHAN, J. ALGINA, and D. B. COULSON, (1978) "Criterion-referenced testing and measurement: a review of technical issues and developments." Rev. of Educ. Research 48: 1-48.

HUBBARD, J. P. (1971) "Objective evaluation of clinical competence," in J. P. Hubbard (ed.) Measuring Medical Education. Philadelphia: Lea & Febiger.

MARQUIS, Y., J. CHAOULLI, G. BORDAGE, S. M. CHABOT, and H. LECLERE (1984) "Patient-management problems as a learning tool for the continuing medical education of general practitioners." Medical Education 18: 117-124.

MARSHALL, J. (1977) "Assessment of problem-solving ability." Medical Education 11: 329-334.

MAZZUCA, S. A., S. J. COHEN, & C. M. CLARK (1981) "Evaluating clinical knowledge across years of training." J. Medical Education 56: 83-90.

McCARTHY, W. H. (1966) "An assessment of the influence of cueing items in objective examinations." J. Medical Education 41: 263-266.

McGUIRE, C. H., L. M. SOLOMON and P. G. BASHOOK (1976) Construction and Use of Written Simulations. New York: Psychological Corporation.

NEWBLE, D. I., R. G. ELMSLIE, and A. BAXTER (1978) "A problem-based criterion-referenced examination of clinical competence." J. Medical Education 53: 720-726.

NEWBLE, D. I., J. HOARE, and A. BAXTER (1982) "Patient management problems: issues of validity." Medical Education 16: 137-142.

NEWBLE, D. I., J. HOARE, and R. G. ELMSLIE (1981) "The validity and reliability of a new examination of the clinical competence of medical students." Medical Education 15: 46-52.

NORMAN, G. R. and J. W. FEIGHTNER (1981) "A comparison of behavior on simulated patients and patient management problems." Medical Education 15: 26.

PAGE, G. G. & D. W. FIELDING (1980) "Performance on PMPs and performance in practice: are they related?" J. Medical Education 55: 529-537.

POPHAM, W. J. (1975) Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall.

SEDLACEK, W. E. and L. W. NATTRESS (1972) "A technique for determining the validity of Patient Management Problems." J. Medical Education 47: 263-266.

SHEPHARD, L. (1980) "Technical issues in minimum competency testing," in D. C. Berliner (ed.) Review of Research in Education, vol. 8. Washington, DC: American Educational Research Association.

WHITE, R. E., B. B. QUIMBY, B. J. SKIPPER and G. D. WEBSTER (1984) "Cost of residents' decisions on actual patients and in simulated encounters." J. Medical Education 59: 833-835.

WOLF, F. M. (1984) "Validity of patient management problems re-examined." Medical Education 18: 222-225.

WOLF, F. M., J. O. WOOLLISCROFT, J. G. CALHOUN, and P. K. HORVATICH (1982) Introduction to Medical Interviewing and History-Taking. Ann Arbor: University of Michigan Medical School, Departments of Internal Medicine and Postgraduate Medicine/Health Professions Education.