Medical problem-solving assessment: a review of methods and instruments

Vu, Nu Viet

How to cite

VU, Nu Viet. Medical problem-solving assessment: a review of methods and instruments. In: Evaluation & the health professions, 1979, vol. 2, n° 3, p. 281–307. doi: 10.1177/016327877900200302

This publication URL: https://archive-ouverte.unige.ch/unige:26644
Publication DOI: 10.1177/016327877900200302

MEDICAL PROBLEM-SOLVING ASSESSMENT
A Review of Methods and Instruments

NU VIET VU
Southern Illinois University
School of Medicine

The purpose of this article is to review and discuss the different methods and instruments created to assess medical problem-solving. Discussions and comparisons of various features of the instruments and suggestions for additional data to be collected were provided in order to insure maximum efficiency in the choice of the instruments.

AUTHOR'S NOTE: This article has benefited substantially from the constructive review of Dr. Terrill A. Mast, Dr. Rosalia Paiva, Dr. David Silber, and Dr. Reed Williams. The author also wishes to thank John Markus for his help in the search of literature and Phil Davis for his suggestions and relentless editing work. Requests for reprints should be sent to Nu Viet Vu, Southern Illinois University School of Medicine, Office of Curriculum Affairs and Educational Resources, P. O. Box 3926, Springfield, IL 62708.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 2 No. 3, September 1979 281-307

One of the main objectives of medical education is to help students develop the ability to solve medical problems. Medical problem-solving incorporates several processes—gathering data, developing and verifying hypotheses, evaluating results, and making final clinical decisions (Vu, 1978). In traditional instruction, students learn the clinical diagnostic process by working with real patients. Performance is assessed through direct observation. The traditional method, however, has limitations. It is often difficult to find the variety of patients necessary to provide students with experences to sufficiently expose them to a wide breadth of diseases. Direct observation has its own inherent problems, especially in obtaining an accurate assessment of students' diagnostic performance. To alleviate these shortcomings, two additional methods of instruction and assessment have been developed: the record-based method and the simulation method. From the basic differences which characterize direct observation, record-based, and simulation methods, several different instruments of assessment have been created for each of those methods.

This article will review, evaluate, and compare these methods of assessment and their respective instruments. An overview of the instruments used with each method, along with some special features which characterize them (purpose, type of assessment, validity, reliability, and specific features) is provided in Table 1. A detailed description of the methods, the assumptions, and their respective instruments will follow.

OBSERVATION-BASED METHOD

The principal problem-solving activity of medical students is to manage an individual patient. Students are expected to gather appropriate data from the history and physical examination, and then derive tentative hypotheses or diagnoses. They should be able to evaluate these tentative diagnoses, collect more data as needed, and arrive at an appropriate final diagnosis. Then they must prescribe treatment and judge its effects on the patient.

TABLE 1
Summary of Methods and Respective Instruments Assessing Medical Problem-Solving

Method	Instrument	Purpose	Type of Assessment	Validity	Reliability	Specific Features
Observation-based method	Rating scale and written report	Assessment and evaluation	Paper and pencil; group administration	NR[a]	NR	-easy preparation and administration; low cost -unstandardized and incomplete measure
Record-based	POMR (1968) Weed	Assessment teaching tool	Paper and pencil; group administration	NR	NR	-easy administration; low cost -no scoring method
	Hospital chart review Williamson(1965) Helfer(1967)	Assessment	Paper and pencil; group administration	Yes[b](content)	No	-easy administration; low cost -chart records may be misleading, unrelated to the diagnosis, and difficult to interpret
	Test for diagnostic skills Rimoldi(1961)	Assessment	Card selection and verbal assessment; group administration	Yes[b](content)	NR	-easy administration; low cost -moderately easy scoring -provides cueing -low fidelity simulation
	Patient management problem	Assessment and evaluation	Paper and pencil; group administration	Yes[b](content)	Yes[b]	-long preparation -easy administration -no feedback - errors cumulative -provides cueing -fixed answer -low fidelity simulation
	Revised patient management problem (University of Illinois) McGuire and Babbott(1977)	Assessment	Paper and pencil; group administration	Yes[b](content) No(concurrent)	Yes	-long preparation -easy administration -difficult scoring -provides feedback - no cumulative errors -flexible answers -provides cueing -low fidelity simulation

TABLE 1 (Continued)

Modified patient management problem (Michigan State Univeristy) Elstein et al. (1978)	Assessment	Paper and pencil; group administration	Yes[b](content) Yes[b](concurrent)	Yes[b]	-long preparation -easy administration -moderately difficult scoring -provides feedback -natural sequence of processes can be observed -provides cueing -low fidelity simulation
Diagnostic management problem Helfer and Slater (1971)	Assessment	Card selection; with paper and pencil; group administration	Yes[b](content)	Yes[b]	-long preparation -easy administration and scoring -provides cueing -low fidelity simulation
Sequential management problem Berner and Tremonti(1975)	Assessment	Paper and pencil; group administration	NR	NR	-long preparation -easy administration and scoring -no cueing -low fidelity simulation
Computer based examination Computerized patient management problem	Assessment and learning tool	Interactive computer; group administration	Yes[b](content)	Yes[b]	-high cost in preparation and administration -standardizes scoring -provides feedback but no cueing -low fidelity simulation
Patient simulation Elstein et al. (1978)	Assessment	Verbal Interaction; Individual administration	Yes[b](construct and content) No[b](discriminant)	No[b]	-long preparation and administration; high cost -difficult scoring -no cueing -high fidelity simulation
Portable patient problem pack Barrows and Tamblyn(1977)	Assessment and learning tool	Cards selection; Individual administration	Yes[b](concurrent)	NR	-long preparation -easy administration and scoring -low fidelity simulation

a. NR = No research available.
b. Results based on studies reviewed in this report.

With the observation method, students are observed on the various aspects that constitute diagnostic problem-solving and evaluated by means of rating scales, written reports, and/or faculty impressions. Barro (1973), in reviewing the major studies on the observational method, focused on three important issues relevant to this method: (a) the dimensions of performance to be included, (b) the definitions and differentiations between "good" and "bad" performance along each dimension, and (c) the weight to be assessed to each category of performance. Based on available data related to these issues, she concluded that more evidence is still needed to decide whether direct observation is a valid and reliable measure of medical problem-solving performance. In addition, this method needs refinement because it is unstandardized and incomplete; because the cognitive skills to be evaluated are generally listed in a global manner and are interpreted differently by different evaluators; and, as Barro (1973) indicates, because the "direct observation approach needs to be validated against outcomes" to prove its validity.

RECORD-BASED METHOD

Problem-solving processes also can be assessed in retrospect from recorded information. Weed (1968) had devised the problem-oriented medical record (POMR) which, according to him, is useful not only for assessment, but also as a teaching device. The POMR "calls for a systematic account of the collection of data, the formulation of problems based on the data, the development of plans and treatment for each problem" (Barro, 1973). In theory, the POMR provides a structure that describes the thought processes of each physician and indirectly reveals the approach to problems that the physician uses. Up to now many articles have been written about POMR, although little of the available literature deals with it as a way to look at the problem-solving process. Several authors (Dinsdale, et al., 1975; Feinstein, 1973; Sandlow et al., 1977; Walker, 1976) have generally agreed that

more research is needed to establish whether POMR is advantageous as either a diagnostic or a problem-solving aid.

Another version of the use of medical records to assess clinical competence has been studied by Scott and Sniderman (1973). In this study, a number of interns dictated their evaluations of patient problems and their investigative and therapeutic plans for the patient. Interns' individual patient assessments as well as the assessments they made with a team were transcribed onto hospital charts; the charts are then evaluated and analyzed by a chief medical resident and attending physician in order to determine the interns' clinical problem-solving success and error. The study failed to prove the effectiveness of this method, however, because the validity was affected by four factors: incompleteness of measured problem-solving processes, competence of evaluators, indirect analysis of performance, and potential influence of the study effect on resident's performance. In addition, although hospital chart review may prove useful in assessing clinical problem-solving, extreme care must be exercised, because written information often does not reflect what actually occurs between the patient and the physician, nor is it necessarily related to the outcome of a diagnosis (Barro, 1973).

A different approach using review of hospital charts to assess physician's performance was described by Williamson (1965) and Helfer (1967). In this approach, a subject was presented with a brief clinical description of a patient followed by a list of possible diagnostic and therapeutic interventions that the subject could select. The problem-solving was evaluated by three scores: (a) an Efficiency Index, which is defined as the percentage of the physician's selections that were helpful; (b) Proficiency Scores, which estimate the quality of product inferred from the percentage of agreement with the criterion group in selecting beneficial and avoiding harmful interventions; and (c) a Competence Index, which indicates percentage of overall agreement with criterion judgment. However, no empirical data were reported to permit assessment of the reliability and validity of the hospital chart review as used in this approach.

Goetz, Folse, and Peters (1976), in using the chart review technique to specifically measure the skills of interpreting information and separating relevant from irrelevant cues, found this technique to be fairly reliable with adequate face and content validity. However, they also found that performance as measured does not generalize from one chart to another; rather, it appears to be task-specific.

In general, the instruments used in both the observation and the record-based methods seem *potentially* useful for assessing problem-solving processes and skills, but refinements are needed before either method will provide accurate and reliable measurement.

SIMULATION METHOD

Besides assessing problem-solving processes and skills through observation and chart review, other instruments have been devised using simulated physician-patient encounters. Several of these instruments are of major interest: the Test for Diagnostic Skills, the various Patient Management Problems (PMP), the Diagnostic Management Problem (DMP), the Sequential Management Problem (SMP), the Simulated Patient (SP), the Computer Based System (CBX), and the Portable-Patient-Problem-Pack (P4). In simulated problems, the subjects have several choices of routes to gain information and with each choice made new alternatives are available. It is assumed that the pattern followed by subjects in making choices and utilizing data within a simulation provides a picture of the problem-solving process.

TEST FOR DIAGNOSTIC SKILLS

One of the earliest instruments used was the Test for Diagnostic Skills devised by Rimoldi (1961). The subject is given specific information about a case to be solved. Additional information is obtained as needed by the subject from removable cards which have a question written on the top edge of the card and the answer

written on the reverse side. Three types of questions can be asked relating to patient interview and history, the physical examination, or laboratory tests. Questions the subject asks are rated in three ways: number of questions asked, relevancy of the questions in rapport to the final diagnosis, and order in which questions are asked. By looking at the order and the type of questions asked, it is implied that the subject's problem-solving process can be assessed. Weitman and Coisman (1975), in using Rimoldi's test, investigated some aspects of the process medical students use in addressing a clinical problem. In general it was found that "the outcome (diagnosis) has significant relationship with some aspects of the problem-solving process (amount and pattern of information employed) as well as the stage of training of the subjects" (Weitman and Coisman, 1975). According to these results, Rimoldi's Test for Diagnostic Skills appeared to measure some aspects of problem-solving ability and possess a degree of content validity. The reliability of this instrument has not been established.

PATIENT MANAGEMENT PROBLEM

Original Patient Management Problem—NBME. In an attempt to assess the different aspects of clinical competence, the National Board of Medical Examiners utilizes Patient Management Problems (PMP) in Part III of the NBME examination. The Board's PMPs are designed to measure nine different areas of clinical competence: history, physical examination, tests to be used, diagnostic acumen, treatment, care implemented, continuing care, physician-patient relationship, and responsibilities as a physician. In these PMPs students are given general information about the patient, results from the physical examination, and data from diagnostic tests. They must study the available information and decide what to do next. After selecting a course of action, they are instructed to turn to a separate answer booket which contains a series of inked blocks. As they remove the ink for their selected choices, the students gain information about the results of the course of action selected. After each selection,

the situation changes, a new problem develops, and new decisions must be made. This step-by-step progression, wherein each step is accompanied by an increment of information, is characteristic of this testing method (which is also called the linear programmed testing). Students proceed in a step-by-step fashion through a sequential unfolding of a series of problems and cannot change any answers once made. Their responses, whether right or wrong, are clearly apparent to the scorer. The total score a student receives is "the number of correct choices made (the number of indicated procedures selected plus the incorrect procedures avoided)" (Barro, 1973).

Concerning the reliability of the PMPs, which is defined as the stability of scores, Hubbard (1971) has shown that PMPs, as devised by the National Board, are fairly reliable.

For the content validity of the PMP, Barro (1973) indicates that the PMP does not assess proportionately the nine areas of clinical competence it assumes to assess. In an attempt to determine the content validity of the PMPs, Schumacher, Burg, and Taylor (1975) compared the performance of the interns and the third-year medical students on PMPs and found that the performance of the former was better than the latter. These results imply that the PMPs on the NBME seem to measure "something that is occurring in direct clinical training . . . that the thing learned is relevant to actual practice" (Barro, 1973). Hubbard (1971) has indicated that the correlations calculated between the subjects' performance on the PMP section of the Part III and the Part II of the NBME ranged from .34 to .48. These coefficients, according to Hubbard (1971), "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 a degree independent of factual knowledge." Comparing the subjects' performance on the PMPs with their performance on Part II of the NBME, and comparing the performance of the interns and the third-year medical students on the PMPs only partially confirm the content validity of the PMPs. For further confirmation, the subjects' performance on the NBME PMPs should also be com-

pared to the direct measure of their performance in a real patient encounter.

Revised Patient Management Problem—University of Illinois. In an attempt to create a more objective and more easily administered test than the PMP as devised by the National Board, McGuire and Babbott (1977) have revised the PMP by modifying its content and procedures. According to Barro (1973), the revised PMP contains four major aspects that distinguish it from the original:

(1) The statement of the problem or case contains less information.

(2) The revised PMP requires a series of sequential and interdependent decisions representing various stages in the management of the patient (branching program).

(3) The feedback of the branching program allows the subject to receive information about the results of each decision as a basis for subsequent action, and allows for different medical approaches and for variation in patient responses appropriate to the approaches. The revised PMP requires the subject to make not a single, correct solution, but to make choices from a large number of strategic routes, several of which may lead to an acceptable result.

(4) The revised PMP differs from the National Board's PMP in its scoring technique.

In order to assess performance on the revised PMP, McGuire and Babbott (1977) have devised five scores: Efficiency, Proficiency, Errors of Omission, Errors of Commission, and a Composite Index of overall competence.

Although McGuire and Babbott's (1977) method of scoring PMPs is valuable in assessing the different aspects of clinical problem-solving, it does have limitations. One major problem derives from the complex system of weighting the items. Because of the weighting system, revised PMPs seem unable to reflect qualitative differences between decisions. For example, a choice which results in the death of the patient is qualitatively different from a choice which only causes discomfort or financial difficulty.

In order to compensate for these limitations, Donnelly (1976) suggested a different method of weighting the items that is both simpler and easier to interpret. He proposed a different computation formula for the Proficiency score which takes into account the relative value of the indicated and contraindicated choices made. This single score proves more consistent with faculty evaluation and also more readily identifies students' deficiencies in clinical problem-solving.

Another problem in the revised PMPs scoring system is described by Marshall (1977). With McGuire's scoring system, respondents who reach a correct confirmed diagnosis with only essential information have the same score as those who take more circuitous pathways and accumulate a mass of data in the process. In order to overcome this deficiency, Marshall (1977) suggested that a maximum mark should be assigned to each individual section in a PMP. The efficient problem solver would then receive all the marks possible in each section, while the respondent who generates superfluous data would not benefit in doing so. This system of scoring, then, allows the efficient performer to be more readily distinguished from the inefficient one. Depending on the PMP users' intentions and priorities, the scoring proposed by Donnelly (1976) or Marshall (1977) can provide them with alternative ways to measure performance.

Although the revised PMPs have been used by some medical schools in order to assess problem-solving and clinical decision-making, few studies have been done to verify their validity and reliability. Content validity can be assessed easily, since the revised PMPs are constructed to include the main components involved in clinical problem-solving. However, it is difficult to determine the PMP's concurrent validity (approximating the performance on PMPs to performance in the real clinical encounter) due to the fact that the PMP format forces respondents to select among available questions rather than initiating their own. Furthermore, the format does not allow subjects to shift back and forth between the activities of history and physical as they see the needs (Goran et al., 1973). Goran et al. (1973) studied the performance of physicians and medical students in a revised

PMP focusing on urinary tract infection (UTI) and compared it to their performance in a real clinical setting. It was found that the subjects were more thorough in their pursuit of a differential diagnosis of UTI in the revised PMP format than they were in a real clinical setting. The performance on the revised PMP, in fact, did not discriminate between the poor, average, and excellent problem solver. The "average" physician did better on the revised PMPs than in actual practice—in fact, those who did best on the revised PMPs did not consistently do best in actual practice. The results from this study, however, should be interpreted and used with caution: First, the data obtained about the clinical encounters were extracted from charts and, as not all findings from an encounter are recorded, results derived from charts may be biased. Second, as the study was unable to assess identical cases for both formats and looked instead at cases comprised of common complaints and diagnoses, comparisons of performance may not be valid. Feightner and Norman (1976), in an attempt to use identical cases in both the revised PMP and simulated patient (SP) formats, have found that performance measured by a revised PMP format is different from that measured using a simulated patient, and that the difference occurred throughout the history, physical, investigation, and management. (Performance in this study was defined in terms of frequency of options selected by subjects.)

In general, the concurrent validity of revised PMP remains unconfirmed. Overall performance on PMPs has not been shown consistent with performance in either real clinical encounters or with simulated patients. As the scores on PMPs derive from algorithms and the scores obtained in a real clinical encounter or with simulated patients do not, further studies might investigate whether this undemonstrated concurrent validity is due to the different systems of scoring used by the various formats.

McGuire and Babbott (1977), in studying the PMP's reliability, had to define reliability as "consistency of measurement" or "the extent to which a particular set of scores in simulated exercises is generalizable to many possible similar tests" (Cattell, 1974; McGuire and Babbott, 1977). Since each of the items within

a PMP is weighted differently, is interdependent, and provides different amounts of feedback to students, the reliability of the PMP could not be "commonly estimated" as other instruments which usually consist of independent items measuring a same trait. McGuire and Babbott (1977) have found that for tests which consist of one fairly lengthy problem, the coefficient of reliability ranges from .75 to .85; for tests which have two or three problems within one medical specialty, the reliability varies between .80 to .90; and, finally, for tests comprised of 10 to 12 problems in different medical specialties, the coefficients range from .85 to .94. Based on the findings, McGuire and Babbott have concluded that the performance measure increased in reliability as the number of problems given to the subjects increased. In other words, the measure of students' performance is more reliable when they are tested on several different problems.

Modified Patient Management Problem—Michigan State University. A third type of Patient Management Problem was devised at Michigan State University by Elstein, Shulman, and Sprafka (1978) and used in the Medical Inquiry Project. In general, it is similar to the one devised by McGuire and Babbott 1977) at the University of Illinois, but since it is used as an observation tool rather than an evaluation instrument, some modifications were made in the format and scoring method. The Modified Patient Management Problem allows two main kinds of observation: (a) the "natural sequence" of processes that a subject would utilize to manage patients and (b) the order in which information is collected. Three scores are calculated instead of five: Efficiency, Thoroughness, and Diagnostic Accuracy.

In order to determine the content and concurrent validity of the modified PMPs, physicians' performance on modified PMPs was compared to their performance on the original PMP format, and to simulated patient format. Content validity for the PMPs appears confirmed, since the overall processes used by the subjects to solve modified PMPs closely resembled those employed in patient simulations. The concurrent validity of the modified PMPs was determined by comparing the means and

standard deviations of Thoroughness, Efficiency, and Accuracy scores on the modified PMP format with those from both the original PMP and the simulated patient formats. The relationship between scores on the modified PMPs and those on the simulated patient proved to be no stronger than the score relationship between original PMPs and simulated patient problems. These results are difficult to interpret, however, because the content of the problems was not identical for the three types of simulations.

The reliability of the modified PMPs was estimated using the Angoff formula on internal consistency (Angoff, 1953). In general, the problems used in the Medical Inquiry Project appear to be internally consistent. Nonetheless, the reliability estimate determined in this study remained relatively weak because the items in each problem were redundant and interdependent, violating the assumptions underlying Angoff's formula, which is based on the independence of the items. Elstein et al. (1978) have suggested that the reliability of the problems would best be determined based on a test-retest measure so that the results can be more generalizable.

DIAGNOSTIC MANAGEMENT PROBLEMS

In a similar attempt to measure the process used by students in solving clinical problems, as well as to provide an easier way to score performance, Helfer and Slater (1971) created a new instrument called the Diagnostic Management Problem (DMP). It is a modified version of both the Rimoldi test and Patient Management Problems. In this instrument the subjects are presented with 96 cards which represent one clinical problem. Each card contains a specific historical fact, a physical finding, or a single laboratory result. The subjects are also told the setting in which they are working and are given a brief description of the case and an index sheet which itemizes the type of information available on each of the numbered cards. The students proceed to work through the problem by selecting as many cards as they desire and in any order they want. They record on an answer

sheet the number of each card selected and the order in which the cards are chosen. Primary and secondary diagnoses are also recorded. Students' scores are based on their performance and compared to defined criteria. Each student receives four scores: Process, Efficiency, Competence, and Diagnostic Scores. These scores are based on the order in which the cards are selected, the total number of cards chosen, and the usefulness of the selected cards.

To determine the reliability of the DMP (defined here as the consistency of performance across problems), Helfer and Slater (1971) asked 19 senior medical students to solve two dissimilar DMPs, and compared their scores on the two tests. The obtained reliability coefficient was .66, although this increased to .80 when students were compared on four tests instead of two.

Robinson and Dinham (1977) attempted to determine the reliability of DMPs by looking at the internal consistency of subscores in two sets of DMPs containing three systematically arranged problems apiece. Each problem was considered as an item. In general, they found that there was high internal consistency for some subscores of the DMPs in each test package and mixed or low internal consistency for the other subscores. In addition, internal consistency of subscores varied by test package, although Process and Omission scores showed a general high correlation. Robinson and Dinham's (1977) findings suggest that the DMPs do not seem to be reliable tests, and, consequently, using them to assess problem-solving processes accurately may be difficult.

Concurrent validity for DMPs was estimated by comparing the score of 42 senior medical students on DMPs with their scores on NBME Part II (which mainly measures factual recall), their scores on PMPs, and their evaluation by faculty. According to Helfer and Slater (1971), if the DMPs did in fact measure the process, the correlation of the scores on DMPs should be low with NBME scores measuring recall, higher with the subjective evaluation by faculty, and higher still with performance on the PMPs. The correlations found were, respectively: .009, .40, and

.60. The authors concluded that DMPs indeed measure not factual recall, but rather the process of problem-solving.

In studying the construct validity of DMPs, Robinson and Dinham (1977) compared students' performance on DMPs before and after a Pediatric/Internal Medicine Clerkship. It was expected that if the tests were valid, they would prove sensitive to prepost clerkship differences in problem-solving skills. In comparing the mean scores of pre-post-clerkship performance, they found significant differences in the scores of Process, Omission, Efficiency, and Diagnosis. On the other hand, when construct validity was defined in terms of the difference in performance among students beginning their first clerkship and those starting their third clerkship, the results were less promising. It was found that more experienced students scored higher than less experienced students on the Process, Omission, and Diagnosis subscores, but not on others. Although the concurrent and construct validity of DMPs are somewhat confirmed in this study, caution is necessary in interpreting the data because the two DMP tests used in Robinson and Dinham's (1977) study were found to be neither reliable nor parallel.

Although the DMPs provide an easier method of scoring and a finer breakdown of the scores when compared to the revised (McGuire and Babbott, 1977) and the modified PMPs (Elstein et al., 1978), all three formats have one main disadvantage in that they all include a cueing aspect in their instruments. All three formats, by listing possible options and alternatives from which students can choose, provide clues for diagnosis that are not readily available in real-life situations. In effect, this characteristic may introduce a biased measurement of performance in that cued situations can provide higher scores of performance than uncued situations, especially for poorer students (McCarthy, 1960). Provision of cues not only provides an artificial clinical situation in which problem-solving performance cannot be validly measured, but also explains why the performance on revised or modified PMPs and on DMPs are not comparable to performance with a simulated patient or in real clinical encounters.

SEQUENTIAL MANAGEMENT PROBLEM

In an attempt to reduce the cueing aspects found in the revised PMPs (Elstein et al., 1978; McGuire and Babbott, 1977), Solomon (Martin, 1975) from the University of Illinois devised a different problem format called the Sequential Management Problem (SMP), in which students have to generate their own questions in order to obtain information. In return, they receive immediate feedback about their request, which allows them to decide the next step to be taken in order to reach a diagnosis. In the SMP format, contrary to the revised PMPs (McGuire and Babbott, 1977) and the DMPs, students also receive corrective feedback which allows them to make fewer cumulative errors as they proceed to their diagnosis. Martin (1975), in studying physician's performance on the revised PMP and SMP formats presenting an identical case, has found that physician's performance (a) is better in history and physical examination in the revised PMP format than in SMP, and (b) is better in the laboratory selection and management in SMP format than revised PMP. This superiority of the laboratory selection and management performance in SMPs could be due to two factors: first, in the SMPs, as the students are provided with corrective feedback, their errors are not cumulative; and second, as revised PMPs provide subjects with alternatives to choose from instead of originating their own questions as in SMPs, the subjects may not be able to assimilate all the given information effectively in order to decide on the appropriate management. Besides this finding, which suggests little equivalence between the SMP and revised PMP formats, no data on the validity and reliability of the SMPs is yet available. SMPs, however, do have several positive features in format and scoring: (a) as each section of the SMP provides a separate score, the areas of strength and weakness are readily detected; and (b) the feedback provided to the students in the SMP also may instruct them toward correct analysis and management (Berner and Tremonti, 1975).

COMPUTER-BASED EXAMINATION AND COMPUTERIZED PATIENT MANAGEMENT PROBLEM

Another patient management simulation that has been developed to measure patient management skills is the Computer-Based Examination (CBX). This system was derived from a project at the University of Wisconsin (Madison) with the support of the National Board of Medical Examiners and the American Board of Internal Medicine. In this system, students operate a standard computer terminal using the CBX model to do a work-up of a patient case. The simulation starts by presenting a patient complaint. Users can then interact with the computer in order to gather information, order tests, and formulate diagnoses. The computer also provides information about the progress of the patient and the test results. Analysis of the users' performance is based on the efficiency with which they handle the various cases and the sequence and degree of efficiency of the tests and therapies they ordered. In addition, users are advised to minimize risk caused to a patient, cost involved, length of stay, time to start corrective therapy, and time spent on a particular patient case. Schultz, Newsom, Entine, Neal, and Friedman (1975) have indicated in a study of residents' performance with the CBX that based on their residents' evaluations of the system, CBX generally reflects real-life performance on actual patients.

Another technique for assessing clinical skills using the computer is the Computerized Patient Management Problem, which attempts to simulate certain components of the physician-patient encounter. In general, Computerized Patient Management Problems (CPMPs) are considered more objective than oral assessment of clinical skills because the subject matter and the scoring are standardized (Grace et al., 1975). In addition, compared to the paper-and-pencil PMPs, the CPMPs prevent students from "retracing" or looking ahead to subsequent items and disallows the possibility of additional choices after leaving a given problem (Schumacher et al., 1975). These characteristics render the CPMP format a closer approximation of the real-life setting; in either format one cannot go back in time to take an

action that already should have been taken, nor can one look in the future to see the results of an action not yet taken.

The CPMP format currently has been used as a complementary testing procedure by the National Board of Medical Examiners and the Royal College of Physicians and Surgeons of Canada. To determine validity, CPMPs are used in conjunction with both a multiple-choice exam (MCQ) and an oral examination as a certification process of the Royal College (Shakun et al., 1976). Students must pass the MCQ and the CPMPs before taking their oral examinations. A factor analysis showed that while orals and MCQ tend to measure basic knowledge, CPMPs measure general components of problem-solving. On the other hand, Schumacher et al. (1975), in comparing linear-written PMPs and CPMPs, have found that by transforming linear-written PMPs into CPMPs, the reliability and validity of CPMPs remain nearly the same. As the results are still limited, further studies are needed to confirm these results and validate the CPMP format.

PATIENT SIMULATION

The Patient Simulation consists of physicians interacting with patients who are trained actors. As in real-world medicine, the physicians do the history and physical exam and decide which data are needed and what laboratory tests are necessary. The Patient Simulation is used in many different ways. Elstein et al. (1978), for example, videotaped the work-ups and encouraged physicians to "think aloud" by giving accounts of their decisions and relating their thoughts at different points in the interview. This kind of introspection technique was used as an attempt to move beyond observable behavior into the thought processes involved in the physicians' problem-solving. In using simulated patients, Elstein et al. (1978) found the technique beneficial in determining, defining, and tracing the processes involved in problem-solving. In addition, construct and content validities of the Simulated Patient were confirmed, but discriminant validity was not (Elstein et al., 1978). In other words, performance with

patient simulations did not provide a clear discrimination between "criterial" and "noncriterial" physicians because individual physicians' performance varied from one case to another. Inconsistencies in performance necessarily affect the reliability of an instrument or the degree of stability with which it measures problem-solving performance. Nonetheless, although the study failed to confirm discriminant validity and reliability in patient simulation instruments, these findings should be interpreted with caution. According to Elstein et al. (1978), inconsistency in performance could have been affected by several factors inherent in the study, such as insufficient variability in the sample, the limited number of problems, or the number of questions used within the problems.

PORTABLE PATIENT PROBLEM PACK—P4

The Portable Patient Problem Pack is a method of simulating a patient's problem in a card deck format devised at McMaster University, Ontario, Canada. Each problem consists of a deck of cards in a variety of colors, as well as slides, printed instructions, and evaluation materials. Colors characterize the type of action taken by the students, such as history, physical examination, investigations, consultations, and treatment or management intervention. The front of each card describes the specific action that the card represents and displays a series of questions that the students should ask. These questions are listed in order to help shape the effectiveness and efficiency of the student's problem-solving skills. Cards for actions such as investigations and consultations may have a number which indicates the time delay involved in obtaining the result of that action in the real clinical situation. The back of each card gives the students answers to questions asked on the front or results of the physical examination or investigations. According to Barrows and Tamblyn (1977), this method allows students to take any action they find appropriate and to see the results of each action before proceeding to the next step. It can be used both as a learning tool and as an evaluation instrument. As a learning tool, for

example, students may stop any time during the problem to read, study, or confer with faculty in order to improve their problem-solving acumen. As an evaluation instrument, the P4 deck assesses students' problem-solving abilities through the pattern of cards accumulated.

The concurrent validity of the Portable Patient Problem Pack has been assessed by comparing performance between the P4 deck and a simulated patient encouter presenting the same problem (Tamblyn and Barrows, 1978). The only significant differences found between the two formats were in the students' consultant and management actions: more actions were taken with the P4 than with the simulated patient. Tamblyn and Barrows (1978) explained that these differences in performance may derive from a confounding factor: some confusion by the students about the use of the P4's time delay and consultant cards. Otherwise, no differences were found between the two formats in the frequency of actions taken on history or examination, the overall outcome, or the process score. The last findings reflect an apparent validity of the Portable Patient Problem Pack and suggest its potential utility for teaching and assessing problem-solving skills. The reliability of the P4 format has not yet been established.

SUMMARY

The various methods of measurement and their respective instruments in the area of clinical problem-solving have been reviewed and discussed. These methods consist of the observation-based method, the record-based method, and the simulation method. Instruments used in the observation-based method rely largely on rating scales and written reports. In the record-based method, the major instruments are the Problem Oriented Medical Records (Weed, 1968) and hospital chart reviews (Helfer, 1967; Williamson, 1965). Instruments utilized in the simulation method are: Test for Diagnostic Skills (Rimoldi, 1961); various Patient Management Problems (Barro, 1973;

Elstein et al., 1978; McGuire and Babbott, 1977); Diagnostic Management Problems (Helfer and Slater, 1971); Sequential Management Problems (Berner and Tremonti, 1975); problems based on computer (Grace et al., 1975; Schultz et al., 1975); Patient Simulations (Elstein et al., 1978); and, finally, the Portable Patient Problem Pack (Tamblyn and Barrows, 1978). Although each of these instruments shares a common format—presenting a patient problem or a case to be solved—they differ in several aspects. As shown in Table 1, each instrument varies in its purpose, format, cost, degree of difficulty in preparation and administration, validity, reliability, and scoring. The variables which characterize each instrument must be considered carefully by potential users to determine which of the instruments is most appropriate for specific educational intentions.

Although the instruments have been devised for specific *purposes,* they can be adjusted according to the users' needs. An instrument created to assess problem-solving can be adapted to be used as an evaluation or a teaching tool. For example, the revised Patient Management Problem (McGuire and Babbott, 1977), which was created to assess performance, also can be adapted to evaluate the strengths and weaknesses of students' problem-solving or to allow them practice in making differential diagnoses.

The *format* of the instrument allows the users to choose individual administration (Patient Simulation, Portable Patient Problem Pack) or group administration (Patient Management Problem, Sequential Management Problem) modes of assessment (paper and pencil, verbal or computerized assessment), provision or nonprovision of cueing and feedback, and degree of fidelity in simulation. As the cost and the degree of difficulty of test administration depend on the format utilized, users must consider these aspects in order to choose the appropriate instruments.

The scoring methods and even the type of scores obtained in each instrument also vary. For example, although the revised PMP (McGuire and Babbott, 1977) and the DMP (Helfer and

Slater, 1971) each have five score indexes, the two instruments have different criteria for measured performance and consequently different types of scores. As a result, it is important for the users to interpret both the score and the performance assessed by the instrument in the context from which they are obtained. Scores obtained by subjects on PMPs and DMPs should be interpreted differently from scores obtained from patient simulations: the latter instrument may be more difficult because it provides no cueing, while a Patient Management Problem does.

As the review demonstrated, *reliability* has not yet been determined for all the instruments. In fact, even for the instruments which have had their reliability calculated, the reliability often remains unconfirmed. Two indexes of reliability are calculated: the internal consistency of the problem and the stability of performance across problems. With traditional methods of calculation, the internal consistency of a problem assumes the independence of the items within it. As most medical problems consist of dependent items, traditional methods to determine the internal consistency are not appropriate. On the other hand, when reliability is measured by consistency of performance across problems, the instruments are generally found to be unreliable because performance appears to be case-specific. In an effort to overcome these problems, Speedie (1976) has suggested that reliability should be redefined and measured according to the general purposes of the investigation. For example, if the investigation is interested in determining general human problem-solving characteristics, it is important to control the experimental conditions to improve the accuracy of the measurement. On the other hand, if an investigation tries to detect individual differences in problem-solving ability, it is more important to have statistical estimates of the accuracy of the measure.

Most of the instruments reviewed have demonstrated some *content validity* in that they do measure the general processes involved in clinical diagnosis. The *construct validity,* however, has not been determined for any of the instruments except for the

Patient Simulations (Elstein et al., 1978). It was found that Patient Simulation problems measure variables that parallel the ones described in theories covering other content domains of problem-solving as well as those in broader theories of cognitive functioning (Elstein et al., 1978). The *concurrent validity* has only been investigated for the revised (Feightner and Norman, 1976; Goran et al., 1973) and modified (Elstein et al., 1978) Patient Management Problems, and the Portable Patient Problem Pack (Tamblyn and Barrows, 1978). To determine their concurrent validity, the performance on revised PMPs was compared with the performance on the same problems in a real clinical setting (Feightner and Norman, 1976), while the performance on modified PMPs and with the Portable Patient Problem Pack were compared with the performance obtained on Patient Simulations (Elstein et al., 1978; Tamblyn and Barrows, 1978). Concurrent validity for revised PMPs remains unconfirmed (Feightner and Norman, 1976; Goran et al., 1973), while validity for modified PMPs and the Portable Patient Problem Pack are only partially confirmed. These findings indicate that performance obtained from low-fidelity simulations (that is, revised PMPs, modified PMPs, and Portable Patient Problem Pack) is not comparable to performance obtained in high-fidelity simulations (Patient Simulation) or in the real clinical setting. This raises an important question for users: What are the advantages in using low-fidelity versus high-fidelity simulation instruments?

From the review it has been demonstrated that the low-fidelity simulations have several advantages over both the high-fidelity simulations and the real clinical setting. These advantages lie in cost, time involved in preparation and administration of the instruments, and the availability of a wider sample of behaviors and diseases to measure. Disadvantages include the loss of certain aspects of reality and lack of the clinical interview setting. However, even these assumed disadvantages have not been verified: for example, it has not yet been empirically determined which features of reality are essential for inclusion in a simulation. It has not been demonstrated how or even whether the unavailability of a verbal interview affects problem-solving

performance or diagnostic outcomes. It is now important to determine which general characteristics of performance distinguish between high- and low-fidelity simulations. Elstein et al. (1978) suggested that when such characteristics are defined, it may be possible to develop a correction formula or adjustment of scores that will enable us to correlate information collected from high-fidelity simulated situations with information from low-fidelity simulations.

In conclusion, available data on the instruments of each method are incomplete, as is evaluation. In order to obtain a complete evaluation of each instrument, future research will need to provide data on validity or reliability, scoring procedures, the evidence of practicality (for example, time to develop and score, amount of required manpower) and, finally, the level of students for whom an instrument is appropriate.

As the task of choosing an appropriate instrument is far more complex than it seems, such complete data along with a careful and systematic approach in considering the different properties of each instrument will be essential to insure the appropriate decision.

REFERENCES

ANGOFF, W. H. (1953) "Test reliability and effective test length." Psychometrika 18: 1-4.

BARRO, A. R. (1973) "Survey and evaluation of approaches to physician performance measurement." J. of Medical Education 48: 1048-1093.

BARROWS, H. S. and R. M. TAMBLYN (1977) "The Portable Patient Problem Pack: a problem-based learning unit." J. of Medical Education 52: 1002-1004.

BERNER, E. S. and L. P. TREMONTI (1975) "Assessment of skills in data-gathering and problem-solving." Proceedings of the 14th Annual Conference on Research in Medical Education, pp. 27-31.

CATTELL, R. B. (1964) "Validity and reliability: a proposed more basic set of concepts." J. of Educ. Psychology 55: 1-22.

DINSDALE, S. M., M. GENT, and G. KLINE (1975) "Problem oriented medical records: their impact on staff communication, attitudes and decision-making." Archives of Physical Medicine and Rehabilitation 56: 269-274.

DONNELLY, M. B. (1976) "Measuring performance on patient management problems." Proceedings of the 15th Annual Conference on Research in Medical Education, pp. 161-164.

ELSTEIN, A. S. (1974) "The medical inquiry project: major findings and implications for medical education." Proceedings of the 13th Annual Conference on Research in Medical Education, pp. 264-266.

ELSTEIN, A. S., L. S. SHULMAN, and S. A. SPRAFKA (1978) Medical Problem-Solving: An Analysis of Clinical Reasoning. Cambridge: Harvard Univ. Press, 1978.

FEIGHTNER, J. W. and G. R. NORMAN (1976) "Concurrent validity of patient management problems by comparison with the clinical encounter." Proceedings of the 15th Annual Conference on Research in Medical Education, pp. 149-160.

FEINSTEIN, A. R. (1973) "Problems of problem-oriented medical records." Annals of Internal Medicine 78: 751-763.

GOETZ, A., J. R. FOLSE, and M. J. PETERS (1976) "Chart review as a measure of clinical competence." Proceedings of the 15th Annual Conference on Medical Education, pp. 45-50.

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

GRACE, M., W. C. TAYLOR, E. N. SKAKUN, and S. FINCHAM (1975) Computerized patient management problem: an alternative examination technique." Proceedings of the 14th Annual Conference on Research in Medical Education, pp. 111-115.

HELFER, R. E. (1971) "Estimating the quality of patient care in a pediatric emergency room." J. of Medical Education 42: 244-248.

——— and C. H. SLATER (1971) "Measuring the process of solving clinical diagnostic problems." British J. of Medical Education 5: 53-55.

HUBBARD, J. P. [ed.] (1971) Measuring Medical Education. Philadelphia: Lea and Febiger.

McCARTHY, W. A. (1960) "An assessment of the influence of cueing items in objective examinations." J. of Medical Education 41: 262-266.

McGUIRE, C. H. and D. BABBOT (1977) "Simulation technique in the measurement of problem-solving skills." J. of Educ. Measurement 4: 1-10.

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

MARTIN, I. C. (1975) "Empirical examination of the sequential management problem for measuring clinical competence." Proceedings of the 14th Annual Conference on Research in Medical Education, pp. 83-88.

RIMOLDI, H.J.A. (1961) "The test of diagnostic skills." J. of Medical Education 36: 73-79.

ROBINSON, S. A. and S. M. DINHAM (1977) "Reliability and validity of simulated problems as measures of change on problem-solving skills." Proceedings of the 16th Annual Conference on Research in Medical Education, pp. 305-309.

SANDLOW, S. J., P. G. BASHOOK, and W. H. HAMMETT (1977) "Is the problem-oriented medical record really being used?" Hospitals 51: 137-138.

SCHULTZ, J. V., R. S. NEWSOM, S. ENTINE, J. NEAL, and R. B. FRIEDMAN (1975) "Resident physician performance on a patient management computer simulation." Proceedings of the 14th Annual Conference on Research in Medical Education, pp. 154-158.

SCHUMACHER, C. F., F. D. BURG, and W. C. TAYLOR (1975) "Computerization of a patient management problems examination to prevent 'retracing.'" British J. of Medical Education 9: 281-285.

SCOTT, H. M. and A. SNIDERMANN (1973) "Evaluation of clinical competence through a study of patient records." J. of Medical Education 48: 832-839.

SHAKUN, E. N., W. C. TAYLOR, and W. OSBALDESTON (1976) "The relationship between computerized patient management problems and other pediatric certifying examinations." Proceedings of the 15th Annual Conference on Research in Medical Education, pp. 167-171.

SPEEDIE, S. M., D. J. TREFFINGER, and J. C. HOUTZ (1976) "Classification and evaluation of problem-solving tasks." Contemporary Educ. Psychology 1: 52-75.

TAMBLYN, R. and H. BARROWS (1978) Evaluation Trial of the P4 System (Problem Based Learning System, Monograph 4). Ontario, Canada: McMaster University.

VU, N. V. (1979) Describing Teaching & Predicting Medical Problem-Solving: A Review (unpublished)

WALKER, H. K. (1976) "The problem-oriented medical system." J. of the Amer. Medical Assoc. 236: 2397-2398.

WEED, L. L. (1968) "Medical records that guide and teach." New England J. of Medicine 278: 593-600.

WEITMAN, M. and F. G. COISMAN (1975) "Medical student pathways to diagnosis." J. of Medical Education 40: 166-179.

WILLIAMSON, J. W. (1965) "Assessing clinical judgment." J. of Medical Education 40: 180-187.