DOCUMENT RESUME

ED 247 321 TM 840 546

AUTHOR Stevenson, Zollie J., Jr.
TITLE Assessment of the Clinical Performance of Medical Students: A Survey of Methods.
PUB DATE 1 May 83
NOTE 30p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Clinical Experience; *Evaluation Methods; *Graduate Medical Students; Higher Education; Observation; Problem Sets; Standardized Tests; *Student Evaluation
IDENTIFIERS Multiple Measures Approach

ABSTRACT

This review of methods used to assess the clinical performance of medical students focuses on four common assessment approaches: (1) the examination developed by the National Board of Medical Examiners (NBME); (2) systematic, multifactor evaluation methods; (3) observation techniques; and (4) problem based methods. Analyzed in conjunction with each approach were reliability and validity data as well as practicality of the assessment approaches. The reliability and validity data are extensive and high for the NBME. The NBME is the least complicated instrument to administer and score. However, it cannot assess client-clinician interactions utilizing live subjects. Reliability and validity data on observation methods are sparse; and where data exists, the coefficients are generally low. Multifaceted evaluation techniques have provided more accurate assessments of student competence, but require more time and more people to administer multiple assessments. A final issue related to the assessment of clinical competence involves the determination of a generally acceptable definition of competence. (Author/BW)

ASSESSMENT OF THE CLINICAL PERFORMANCE OF MEDICAL STUDENTS: A SURVEY OF METHODS

Zollie J. Stevenson, Jr.,
Graduate Fellow,
Office of Research and Development for
Education in the Health Professions
University of North Carolina Medical School
Chapel Hill, N.C. 27514

May 1, 1983

Abstract

This review of methods used to assess the clinical performance of medical students focuses on four common assessment approaches: 1. The examination developed by the National Board of Medical Examiners (NBME), 2. Systematic, multifactor evaluation methods, 3. Observation techniques, and 4. Problem based methods. Analyzed in conjunction with the approaches review were reliability and validity data as well as practicality of the assessment approaches. The conclusion highlights strengths and weaknesses in commonly used clinical assessment approaches and summarizes related measurement issues.

ASSESSMENT OF THE CLINICAL PERFORMANCE OF MEDICAL STUDENTS: A SURVEY OF METHODS

The traditional method for assessing medical students' clinical competence is through oral and/or written examinations developed by medical faculty. The oral examination questions are presented to individual medical students by a panel of medical faculty members at patients' bedsides or in conference room settings. The questions are to reflect students' clinical experiences. Responses are rated by an examining panel to determine whether students passed or failed. Oral examinations proved to be unreliable measures of clinical competency due to interrater scoring differences, the subjective nature of the assessment procedures, and the problem of defining clinical competence (Levine & McGuire, 1970). Despite the reliability problems, the oral examination remains the major method for evaluating clinical competence in Great Britain, Australia and Canada. The Canadian College of Family Physicians have improved the reliability of the oral examination in Canada by defining minimal clinical competency standards and constructing an objective problem solving examination (Van Wart, 1974). In many countries, the oral examination is employed as one segment of multiple clinical assessment modes. In the United States, the oral examination was discontinued as a subtest of the National Board of Medical Examiners (NBME) because of reliability problems (Hubbard, 1971).

Originally, written clinical examinations were teacher developed and were often unreliable and invalid measures of clinical competence.

Current[illegible] objective based written standardized examination has becom[illegible] method for assessing clinical competence, but reliabili[illegible]tive based and standard examinations have improved due to six [illegible]rs: 1) examination questions reflect predetermined objectives; 2) minimal levels of competency are stated; 3) many tests have been validated (content and predictive validity studies); 4) the reliability of written examinations has been studied; 5) objective test items have eliminated most scoring problems (e.g., interater differences); and 6) competency is based on the extent that performance matches the objectives.

However, the use of written standardized tests has presented another problem. The uniqueness of medical school clinical programs and the richness of individual student experiences may not be tapped by written standardized examinations. Many schools employ the written examination in conjunction with other assessment methods to paint a clearer picture of medical students' clinical competence.

This paper will review: 1) approaches used to define clinical competency of medical studnets; 2) methods for assessing clinical competency; and 3) outcomes of the assessment approaches. The latter will focus on reliability, validity and practicality of the clinical competency assessment methods.

Definitions of Clinical Competence

Rarely have explicit statements been presented of what the undergraduate medical student is expected to perform and at what level of proficiency. Some performance goals have been reflected in objective

based written examination items but many studies suggest that the correlation between the written examination grades and actual clinical performance is usually small and sometimes inverse (Wingard & Williamson, 1973). Instructional designers have recommended using behavioral objectives as means for refining the definition of clinical competence. The use of behavioral objectives has only recently begun to receive widespread acceptance in localized test development.

A different approach involves the analysis of tasks performed by clinicians that are assessed through observation. The tasks are then classified. Task analysis resulted in the development of standards to assess students during the clinical years (Adams & Mendenhall, 1974). Two methods of developing definitions of clinical competence emerged from this approach: 1) identification of elements leading to satisfactory performance; and 2) measurement of performance outcomes regarding patient care.

For classifying clinical competence, the critical incident technique (Flanagan, 1954) has been the most widely accepted approach. Incidents of good and bad performances are identified and classified. Hubbard et al. (1965) identified nine major categories of clinical competence: History, Physical Examination, Tests and Procedures, Diagnostic Acumen, Treatment, Judgment and Skill in Implementing Care, Continuing Care, Physician-Patient Relations and Responsibilities as Physician. Each of the nine categories were defined as operational tasks. With the critical incident technique, classification results from observed outcomes of performances affecting patient care.

Since clinical competence definitions are dependent on observed

outcomes and behaviors, the complexity of observing and defining the constructs and activities to be measured is difficult and complex. Therefore, clinical competency measures are often imprecise.

Methods for Assessing Clinical Competence

Performance assessment is generally measured by analyzing processes to solve problems or by analyzing products or outcomes that result from solutions. Medical student evaluation is conducted entirely on process measures. Since medical students are supervised in their care of patients and thus do not assume direct responsibility for treatment, direct responsibility is a necessity for performance assessment to be based on a product or "production" mode of assessment. Thus, the process mode is the method of evaluating clinical competence.

The review by Wingard and Williamson (1973) indicated that little or no correlations existed between process measures (grades) and future performance. Thus the impetus has arisen for the development of test procedures with predictive validity that improve the product - competing medical school.

A variety of methods exist for the assessment of clinical competence. Four common approaches for measuring clinical competence will be reviewed: 1) the National Board of Medical Examiners examination; 2) Systematic, multifactor evaluation methods; 3) observation techniques; and 4) problem based methods.

National Board of Medical Examiners

The examinations of the National Board of Medical Examiners consist of three parts: 1) Part One, Preclinical Sciences (first two years); 2) Part Two, Clinical Sciences (third and fourth years); and 3) Part Three, Clinical Competence (internship or residency). The Preclinical and Clinical Sciences examinations have been established as highly reliable measures of medical knowledge and a candidate's ability to apply knowledge to the problem at hand (Cowles & Hubbard, 1954; Hubbard & Cowles, 1954). Part Two has yielded lower reliability coefficients than Part One. Reasons cited for the lower reliability of Part Two when compared with Part One are: 1) the increased complexity of Part Two subjects when compared to Part One subjects; 2) the variability of the methods for grading students (resulting in the lower reliability of instructor ratings); and 3) the homogeneity of clinical year students. Statistical studies yielded evidence that NBME, Parts One and Two, generally: correlated more highly with independent estimates of student proficiency by instructors, demonstrated a reliability of measurement more adequate for precise grading, and differentiated among the candidates due to the score distribution (Cowles & Hubbard, 1954).

Before 1961, Part Three of the NBME was usually conducted as a bedside, oral examination (Hubbard, Levit, et al., 1965). A case history and physical examination were taken for a patient. Then the M.D. was questioned by an examiner who was not familiar with the patient. The examiner would then use the patients' chart to develop an examination in the form of a quiz session. The procedure would

then be repeated for the same M.D. with a different patient and examiner. Frequently, the inter-scorer reliability and evaluation from one examination to the next was low or negative. Three variables impacted on the bedside evaluations: the candidate, the patient and the examiner.

Hubbard, Levit, et al. (1965) reviewed the new techniques employed by NBME to validate the clinical competence measure, Part Three. Clinical competence as defined was based on feedback from questionnaires and interviews secured from interns citing incidents of clinical performance. Nine areas were considered in defining clinical competence: history, physical examination, tests and procedures, diagnostic acumen, treatment, judgment and skill in implementing care, continuing care, physician-patient relations, and responsibilities as a physician. Subcategories, defined in behavioral terms, existed within the nine categories. Clinical competence was determined to be best measured by developing Part Three using motion pictures of carefully selected patients, a section calling for the interpretation of presented clinical data (graphic and pictorial form) and programmed testing is used. Thus the patient variable was controlled by standardizing the patient experience viewed for assessment by the students. The questions posed were asked with objective responses as solutions to problems presented (e.g., clinical data, patient motion pictures, etc.). Thus, the resulting examination was a more objective measure.

Test analyses yielded Kuder-Richardson reliability coefficients (for two equivalent forms) to be 0.83 and 0.87. Since many critical incident problems were included in the NBME, Part Three, the test was

stated to have high content validity. No predictive validity data was available in the Hubbard, Levit, et al. study. However, correlations of the NBME Part Three examination with NBME Part Two yielded correlation coefficients ranging from 0.30 to 0.65. The coefficients provided some indication that the results were fairly independent of each other.

Hallock, Christenson, et al. (1977) reported data on the clinical performances of medical students in three and four year medical curricula in five category areas: 1) fund of knowledge; 2) medical skills; 3) problem solving; 4) professional standards; and 5) reliability of the student in performing his/her duties. A five point grading system was used by the faculty to rate student performances in the five areas. The results consisted of assigned points on the five categories by faculty members and NBME scores on the pediatrics, medicine, and obstetrics/gynecology tests. Data analysis indicated that the four year students scores on the NBME correlated most highly on the fund of knowledge category (they also had higher NBME scores). Three year students scored higher on problem solving and professional standards. The Hallock, Christenson, et al. study provided some evidence of the predictive validity of the NBME, Part Three.

The NBME, Part Three, has demonstrated high reliability, high content validity, godo predictive validity and high practicality in terms of administrative ease.

Systematic, Objective Evaluation Methods

A variety of multfaceted approaches for assessing clinical com-

petence have been developed. Most of the development occurred during the late 1960's and the 1970's. The multifaceted assessment approach attempted to measure student performance and competence in the clinical setting. Direct observation, oral and written examinations, each adding to the total measure of clinical competence, were common approaches for assessing clinical competence. In many instances, the NBME was used as one of the written examinations. The goal of the approaches reviewed in this section was to add objectivity, variability and structure to the assessment process so that measurements of competence would be more reliable and valid.

Reviewed in this section are systematic, multifaceted approaches developed by Geertsma and Chapman (1967), Graham (1971), Printen, Chappel and Whitney (1973), O'Donohue and Wergin (1978) and Sheehan, et al. (1980).

Geertsma and Chapman (1967) studied the system of evaluating student performance implemented at the University of Kansas which attempted to measure eleven dimensions: 1) fund of information; 2) comprehension; 3) problem solving; 4) reliability; 5) application; 6) judgment; 7) originality; 8) rapport with patients; 9) poise; 10) ethical standards; and 11) likability. Four additional dimensions were added to aid in the preparation of recommendations and summary reports of student progress: 1) probable success as a student; 2) probable success as a physician; 3) acceptability as a graduate student or house officer; and 4) overall performance.

The total performance dimension could override an unsatisfactory rating on one of the first 11 dimensions or could serve to offset

superior ratings. A three category descriptive rating scale (unsatisfactory-superior) was employed by the instructor of each student's major course to evaluate student performance. In some departments at the University of Kansas, faculty met collectively to evaluate their students. In others, faculty members handed in evaluations and a consensus was reached in the evaluation of each student's work in the course. The dimensions were printed on small cards which contain spaces for narrative information. Analysis of the data indicated that the ratings of the dimensions were highly interrelated with two factors being identified: general cognitive factors and a noncognitive factor centering on ethical standards. Instructors tended to give unsatisfactory ratings on cognitive dimensions and superior ratings on noncognitive dimensions (superior ratings are reported more frequently than unsatisfactory ratings). Since the dimensions were determined a priori, the method suggested that the evaluation dimensions be revised so as to provide operational guidelines for each dimension derived.

Graham (1971) attempted to define behavior expected in a clinical clerkship and developed a method of reporting such performance. The evaluation form for clinical competence has nine sections: 1) attainment of global objectives; 2) descriptive checklist; 3) clinical performance checklist; 4) narrative; 5) suggestions/ comments; 6) career choice recommended; 7) degree of change; 8) other comments; and 9) final evaluation.

The evaluation method is very time consuming (due to its detail) and asks questions that sometimes cannot be answered due to the lack

of familiarity with students (on the part of instructors). At the beginning of the clerkship, students evaluate themselves using the evaluation form which was used as a part of the departmental summary. Faculty, preceptors and staff involved with each student's program also received copies of the evaluation forms for their comments (at the end of the term). The forms served as the basis for evaluating student clinical competence. The report is perused by the student and discussed with the undergraduate coordinator, with emphasis on weaknesses and differences of opinion as well as strengths.

Printen, Chappell and Whitney (1973) implemented a comprehensive, objective evaluation process to assess the clinical performance of junior medical students based on an oral examination, a written examination, clinical performance and psychomotor skills. Their system considered behavioral characteristics, mastery of cognitive material, and performance of psychomotor skills, and culminated in the development of a student profile to provide student feedback and objective evidence of student performance and course evaluation data. Oral examinations were held weekly in small groups (two to four students and the instructor) and focused on the cognitive objectives. Clinical evaluation was based on ratings by at least one faculty member, one resident and one intern, on 10 clinical performance variables previously rated by the surgery faculty. Significant rater differences were investigated thoroughly by the clerkship director. Psychomotor skills were assessed by having the student perform certain tasks and then graded on a pass-fail basis by a resident or member of the surgical staff. The written examination was developed around

the departmental cognitive objectives and focused on patient oriented questions in clinical problem situations. The data were analyzed by computer based on predetermined weights and resulted in a student evaluation profile. The Printer et al. method eliminated some of the subjectivity from the evaluation of students' clinical performance. The authors considered their greatest contributions to evaluation to be the structuring and ordering of clinical performance characteristics on a weighted basis (provided guidelines for assessing effective and ineffective clinical performance).

O'Donohue and Wergin (1978) developed a proficiency assessment process to evaluate the performance of medical students during a clerkship in internal medicine employing preceptor evaluations of on-the-job performance as well as independent written and oral examinations. Preceptor evaluations consisted of ratings, on a four point scale, using standardized evaluation forms. Every student was evaluated by at least one preceptor who then submitted a separate evaluation form. Written examinations were developed based on questions submitted by faculty in each of the clinical divisions in the department of medicine. The questions were mostly of the multiple-choice variety. Thirty minute oral exams were given by two faculty members who had not served as preceptors for individual students every three months.

The examiners were trained in oral examination techniques. They were also presented with a listing of each student's patients and diagnoses for use by the examiner. Each examiner provided individual scores and then jointly decided on an oral exam score.

Final grades were determined by a clerkship committee: the clinical was given a weight of 66%, written and oral examinations, 17% each. Reflected was the opinion of the committee that clinical ratings should carry the most weight in the determination of a final grade. The results of the study indicated that between 10 and 70 percent of the variances in clinical ratings was due to situational variables in the performance of individual students and rater error. Ceiling effect was cited as a possible contributory factor to the low inter-rater correlations. The oral examinations had high reliability (.754) due to the nature of the examination (demonstration of one sample of behavior, and lack of correlation with other measures). Intercorrelations among the three raters of students indicated small intercorrelations (the examinations appear to contribute different kinds of data about student knowledge and competence). The study concluded that neither the oral nor written examinations correlated highly with performance assessment and that considerable intrastudent error existed.

Sheehan, et al. (1980) studied the role of moral judgment in predicting clinical performance. Moral reasoning was assessed by the Defining Issues Test and the Moral Judgment Interview. Clinical performance was assessed by a scale which measured eighteen performance characteristics covering medical knowledge, task organization and interpersonal relations. The results indicated that moral reasoning is a predictor of clinical performance. High moral reasoning appears to exclude the possibility of poor performance. The very highest level of clinical performance appears never to be

reached by those at the lowest level of moral thought. The subjects were residents.

The systematic, objective approaches for assessing clinical competence have in common good content validity, fairly low inter-scorer reliability, no predictive validity or test reliability data. All of the studies defined a conceptualization of clinical competence and structured their assessments based on their definition of competence. As a result, content validity appeared to be high (Geertsma & Chapman, 1967; Graham, 1971; Printen, et al., 1973; O'Donohue & Wergin, 1978; and Sheehan, et al., 1980). Inter-scorer reliabilities ranged from poor to low (Geertsma & Chapman, 1967; Printen, et al., 1973, and O'Donohue & Wergin, 1978).

The systematic, objective assessment approaches require great amounts of time and manpower to administer. Thus, the systematic assessment approach is not the most practical of the four assessment modes under review (Geertsma & Chapman, 1967; Graham, 1971; Printen, et al., 1973). The Sheehan, et al. study (1980) provides some evidence of predictive validity, but moral judgment is related to performance characteristics rather than to examination measures.

Systematic, objective approaches for evaluating clinical competence need further development before their use as reliable and valid measures can be documented. The lack of practicality will be an issue as long as the length of the assessment process and the number of people involved in the process remains as stated in the studies.

Observation Methods

Direct observation by staff members of clinical student performance is a popular evaluation method. The clinical student is usually observed at the patient bedside performing routine tasks such as history taking, rapport building with the patient, physical examination and data synthesis. Reviewed herein are studies by Hinz (1966), Hess (1969), Oaks et al. (1969) and Turner, et al. (1972) employing techniques such as videotaped observation and student-preceptor bedside observation. Rating scales for assessing clinical observation are also discussed.

Hinz (1966) described the development of a method of direct observation of students concerned primarily with performance in history taking and physical examination. Hinz devised a study to examine the following: 1) to determine whether teaching is improved by having the instructor observe at the bedside during the student's case writing; 2) to develop more objective criteria for performance in the case method (to establish quantitative as well as qualitative descriptions of performance); 3) to determine whether direct observation makes apparent aspects of student performance that are not otherwise apparent; and 4) to determine the following effects of direct observation on faculty and students: a) effect on the patient-doctor relationship of having an observer at bedside; b) the reaction of students to being observed; c) the cost to faculty in time; and d) the impact of the faculty on student performance. Components of the patient examination were compiled from listings provided by a group of interns (physical examination, interview and

the organization and synthesis of data). All items were tested for value in meeting patient examinations and for observability. A group of internists and psychiatrists observed a group of volunteer fourth year medical students during patient work-ups. They found that: 1) untrained raters yield inconsistent ratings; and 2) students regarded the experience as an opportunity for tutorial aid in history taking and physical examination. Items were categoried according to portions of the patient examination with particular attention on the content of the illness and the method for securing the history. The items were general and a comprehensive assessment applying to any case could be developed. Fifty items were included in the rating scale with sufficient space for notations. Raters were trained using videotaped medical students conducting patient examinations. The pilot study consisted of raters sitting at bedside as a student did a complete work-up of the patient. Afterwards, the student summarized his findings and presented them to the rater and they discussed the case. After discussing the case, the rater used the recorded observations as the basis for reviewing the student's performance. A great deal of interrater inconsistency was found to exist. The pilot study aided in the development of standards of performance and in enhancing the quality of student skills in subsequent weeks but quantitative limitations existed (a need to weight items, etc.). The rho values of rank order correlations of like pairs of raters ranged from .55 to .79 (like = both from the tapes or live). Rho values were low for "unlike" raters (.42 - .58). Thus live and taped observations were not rated in the same fashion. For interviews, observers recorded an overall grade

indicating whether the interview was good, fair or poor. The rho value for the correlations between score and grade was .88. Direct observation is a potentially useful tool for qualitative evaluation of student performances, especially when that evaluation is used for instruction of the individual student. For quantitative purposes, direct observation has limited use since raters differed significantly in their view of various components of the tasks, and because adequate reliability has not been achieved (due to the inability to structure an adequate test of reliability).

Hess (1969) studied the reliability of two rating scales based on a behavioral definition of skill in evaluating student skills in relating to patients. Format A required the raters to classify single units of student behavior (an uninterrupted, purposeful action by the students) under one or more of the 11 categories. Students were videotaped and their interviewing skills were rated. More traditional Format B consisted of a series of statements which described various effective and ineffective types of observable student actions. Each student was videotaped and their performance evaluated on a 10-point continuum. Rating scores from Form A (the interrater analysis) were more reliable than the scores from Form B (Overall A = 0.92; Overall B = 0.66). The importance of the design of the rating instrument in enabling humans to function as reliable data recording instrument was noted. A rating system which facilitated discrete judgment proved to be more reliable than the instrument requiring fewer but more global judgments. Hess concluded that the interaction analysis format for assessing learning provides a much clearer picture of each student's

interview performance than did Format B and provides more clarity and precision of measurement.

Oakes, Scheinok and Husted (1969) studied an objective rating scale used to assess student performance in a clinical clerkship. The scale assessed clinical clerkship performance in 11 attributes: appearance, deportment, maturity, cooperation, scholastic ability, student effort, interest in service, responsibility, professional competence, interpersonal relations and chart neatness/promptness. Students were rated by preceptors using a four-point descriptive scale (poor-excellent).

An overall numerical rating (ranging from 65-100) was also noted by the preceptor on the card. Objectivity was facilitated by providing the preceptor with a three-page form listing descriptions of the 11 attributes. The descriptions served as guidelines for the overall rating and facilitated an accurate estimate of overall clinical ability because of the need for the preceptor to examine individual components of the student's attitude and performance. This study measured: the reliability of the preceptors awarding objective grades compared with overall grades, the reliability of preceptors' ratings depending on academic rank and the percentage of mismatched grades. The study concluded that preceptors (almost all of them) did give a failing grade when failure was indicated, that 42.7% of the preceptors' grades differed from objective grades by more than 3% (but only 4% differed from objective grades by more than 10%) and that associate/assistant professors and instructors were more reliable in grading clinical performance (fewer mismatches) than were residents and full professors.

Turner, et al. (1972) questioned whether clinical competence could be evaluated by observing the performance of individuals in the patient care situation. Student clinical performance was videotaped and rated by specially trained pediatric residents. The researchers wanted to know if good clinical performance could be differentiated from poor clinical performance. Hess's (1969) method for assessing interpersonal and communication skills was used, using two approaches for the assessment of each variable: 1) a tally of the specific acts which were predefined as contributing to the variables in question; and 2) global ratings. The data indicated that variables used to evaluate clinical performance can be better evaluated through tabulation of specific acts as opposed to global judgments (the form in which the variables are expressed affects reliability). Trained raters agreed on many, but not all, physical examination procedures performed by students. Agreement among the professionals was important in the determination of variables that represent competence (a priori quality standards are poor indicators).

As with the systematic, objective evaluation approaches, direct observation has limited and generally low reliability and validity data. Scorer reliability problems were reported (Hinz, 1966; Hess, 1969; Oaks, et al. 1969; and Turner, et al., 1972). In one study, high predictive validity was indicated when observation scores were compared with grades (Hinz, 1966), but no other predictive validity was indicated in the remaining studies. Reliability coefficients for rating scales which facilitated the formulation of discrete judgments were higher than scales that utilized global judgments (Hess,

1969; Turner, et al., 1972).

The data in support of test construction yielded information which supported the use of observational techniques for improving student clinical performance in a qualitative sense.

The use of rating scales and observation for evaluation clinical competence is a time consuming effort. Rating scales fail to capture all important facets of a student's clinical competence and the raters may prejudice the outcomes of an evaluation by either being too familiar or too unfamiliar with students being observed and rated. Observation is rarely an objective method of assessment.

Problem Based Approaches

The Problem Based Examination approach focuses on defining events likely to be experienced by clinical students and basing assessment of clinical competence based on how the student solves the problem. Studies conducted by Harden, et al. (1975), Newble, et al. (1978), Harden and Gleesen (1979) and Newble, et al. (1981) are reviewed.

Harden, Stevenson, Downie and Wilson (1975) introduced a structured clinical examination requiring students to rotate from one station to another in a hospital ward with various tasks assigned at each station (e.g., station one, carry out some aspect of a physical examination, station two, answer multiple choice questions on the physical examination). The cueing effect that usually exists in multiple choice examinations was minimized because the students cannot go back to check omissions in their actions and thus resulted

in a fairly cue free examination.

The structured examination setting allowed variables and examination complexity to be controlled, aims could be more clearly defined and more of the student's knowledge tested. Thus the examination was more objective and a marking strategy could be decided in advance. The examination resulted in improved feedback to students and staff. Analysis of examination results indicated that poor clinical performance was due to: 1) all around inadequacy; 2) deficiency in some aspect; and 3) deficiency in specific subject areas. A study was conducted grouping traditional clinical examination and objective clinical observation with written examination scores. The traditional scores correlated 0.17 with the written examinations while the objective clinical evaluation scores correlated 0.63 with the written. This method allowed for more control over the testing situation and complexity of the material.

Newble, Elmslie and Baxter (1978) developed a patient problem based method for assessing clinical competence in specific areas. A listing of problems likely to be experienced by interns was derived by a consensus process using a wide selection of clinical teachers. A specialist was asked to develop a patient problem blueprint in such a manner as to make it fit the scope of interns' experience. Interns, residents, etc. reacted to the blueprint. The blueprint was expanded to require more detailed knowledge in key areas. The expanded problem blueprints became the basis for selecting appropriate test methods and for the construction of test items. The criterion was not defined in precise behavioral terms but the problem

blueprint provided precision to test construction. Open ended and multiple choice examination questions were developed for each of 12 expanded test blueprints. Students circulated among examination stations. The examination was administered to senior and junior medical students as well as selected residents and interns (the latter provided criterion levels of performance). The number of participants volunteering time to the task indicated that the new approach was acceptable to faculty members and students. Sixty three percent of the students felt that a mixture of multiple choice and free response questions were appropriate for the final examination, but 84% felt that the free response items gave a more accurate assessment of their ability. The students rated the content test as being of high (47%) or moderate (53%) clinical relevance. Ninety five percent of the students indicated that the practical section contributed to a more accurate assessment of their competence than the traditional clinical examination. The practical section content was rated as either highly (74%) or moderately (26%) relevant. This approach was considered to be practical and feasible to administer.

Harden and Gleesen (1979) discussed a procedure designed to assess clinical competence at the bedside employing the objective-structured clinical examination (OSCE). The OSCE separated competence areas into various assessed components. Each component serves as an objective for each station in the exam. This method paralleled that outlined by Harden, et al. (1975), but provides a detailed method for implementing the procedure. No validity or reliability data were provided.

Newble, Hoare and Elmslie (1981) provided validity and reliability data for the problem based CRT of clinical competence. The results demonstrated that the examinations have a high level of content validity (as assessed by teaching staff and students) and showed some evidence for construct validity. Ninety two and a half percent of students felt the content of the test was of high or moderate relevance. Ninety five percent rated the clinical in a similar fashion. Satisfactory levels of internal consistency were established for the whole test. Marker validity was satisfactory on all test sections except those requiring examinations to rate practical skills. Prediction and concurrent validity data could not be accurately secured due to inconsistency in resident and intern's scoring. The test correlated highly with combined marks in medicine and surgery ($r = 0.62$, p 0.01) with a similar level of correlation existing for the new examination and subsections of the final examination (Medicine $r = 0.54$, Surgery $r = 0.62$). The new examination written component was more highly correlated with the final examination ($r = 0.54$) than the practical component ($r = 0.11$). Scorer reliability for the free response section of the examination was very high (0.95). Reliability in the stations whose students were rated ranged from 0.25 - 0.77.

Reliability data have been reported for the problem based assessment approach (Newble, et al., 1978; and Newble, et al., 1981). Predictive validity data ranging from 0.17 - 0.62 have been reported for respective problem based examinations when compared with other written examinations (Harden, et al., 1978; and Newble, et al., 1981).

The examinations involved students performing specific tasks as they moved through a variety of stations in the clinical setting. This practice is time consuming in the amount of time required to rotate through all of the stations. Scorer reliability was low to good in the one example cited (Newble, et al., 1981).

Conclusions

The literature reviewed provides a summative overview of the methodologies for assessing the competency of medical student clinicians. Of particular interest is the reliability and validity data pertaining to the four approaches.

The reliability and validity data are extensive and high for the NBME examination as an assessment instrument. The NBME is the least complicated instrument to administer and score. The review could have ended with the discussion of the NBME if the medical profession was interested in only what is practical. Measurement problems do exist: 1) the NBME is a standardized, norm-referenced measure; and 2) the NBME is an external examinations used to measure in a nonstandard setting. As a standardized, norm-referenced measure, the NBME assesses a sample of behaviors that may reflect competence. Subtle and situational information about student competence can not be adequately assessed by the NBME. A major gap is the inability to assess client-clinician interactions utilizing live subjects. Pass marks for the NBME are low which may indicate that the NBME is imprecise. The pass mark for the NBME, Part III, was 290 nationally in 1981 (800 is the maximum score). The University of North Carolina Medical School

pass mark was 320 in 1980.

Another issue is the use of an external examination to assess performance in settings with variable curricula and student populations. Wile (1978) provided evidence that the NBME was not a relevant measure for student success at a midwestern medical school as was an objective based examination.

Reliability and validity data on observation methods are sparse. Where the data exists, the coefficients are generally low. Most of the reliability data pertain to scorer reliability. The literature concluded that scorer reliability coefficients are low when observation techniques and rating scales have been used as assessment instruments. The recent efforts toward standardizing observation rating scales has slightly reduced scorer inconsistency (Newble, 1976).

Multifaceted evaluation techniques have provided more accurate assessments of student competence. One multifaceted evaluation strategy involves utilizing a written and clinical observation measure (equally weighted). The NBME or an objective based, teacher constructed examination usually serves as the written measure. Problems associated with the multifaceted approach include the length of time and number of people required to administer multiple assessments. The data indicate that multifaceted tests can be reliable and valid measures, though not the most practical measures.

A final issue related to the assessment of clinical competence involves the determination of a generally acceptable definition of competence which can be utilized by medical examination committees. Clinical assessment definitions have focused on diverse situational

and behavioral definitions of competence: 1) clinician-patient bedside behavior versus patient management problem; and 2) evaluation of students based on a description of character traits versus evaluation of objective measures of clinical performance (Newble, 1976). Definitions of clinical competence occur in medical schools based on a consensus of opinion. Those definitions are reflected in the assessment methods. Multiple measures of competence will likely be preferred over solitary measures when the issue of definition has been resolved.

The use of multiple measures increases the chances of securing an accurate and sensitive evaluation of clinical students. A balance between objective and subjective measures in one evaluative instrument does not exist. Computer technology has the potential for revolutionizing the process of evaluating medical student clinical competence combining the objective with the subjective while eliminating scorer inconsistency.

Bibliography

Adams, P.H. & Mendenhall, R.C. "Profile of the cardiologist: Training and manpower requirements for the specialist in adult human cardiology." American Journal of Cardiology, 1974, 34, 389-400.

Cowles, J.T. & Hubbard, J.P. "Report from NBME on validity and reliability of new objective tests." Journal of Medical Education, 1954, 29(6), 30-34.

Flanagan, J.C. "The critical incident technique." Psychological Bulletin, 1954, 51, 327-33.

Geertsma, R.H. & Chapman, J.E. "The evaluation of medical students." Journal of Medical Education, 1967, 42, 938-42.

Graham, J.R. "Systematic evaluation of clinical competence." Journal of Medical Education, 1971, 46, 625-29.

Hallock, J.A., Christensen, J.A., Denker, M.W., Hochberg, C.H., Trudeau, W.L. & Williams, J.W. "A comparison of the clinical performance of students in three-and-four-year curricula." Journal of Medical Education, 1977, 52, 658-63.

Harden, R. McG. & Gleeson, F.A. "Assessment of clinical competence using an objective structured clinical examination." Medical Education, 1979, 13, 41-54.

Harden, R. McG., Stevenson, M., Downie, W.W. & Wilson, G.M. "Assessment of clinical competence using an objective structured examination." British Medical Journal, 1975, 1, 447-51.

Hess, J.W. "A comparison of methods for evaluating medical student skill in relating to patients." Journal of Medical Education, 1969, 44, 934-38.

Hinz, C.F. "Direct observation as a means of teaching and evaluating clinical skills." Journal of Medical Education, 1966, 41, 150-61.

Hubbard, J.P. Measuring Medical Education. Lea & Febiger: Philadelphia, 1971.

Hubbard, J.P. & Cowles, J.T. "Comparative study of student performance in medical schools using National Board exams." Journal of Medical Education, 1954, 29(7), 27-37.

Hubbard, J.P., Levit, E.J., Schumacher, C.F. & Schnabel, T.C. "An objective evaluation of clinical competence." New England Journal of Medicine, 1965, 272, 1321-28.

Levine, H.G. & McGuire, C. "The validity and reliability of oral examinations in assessing cognitive skills in medicine." Journal of Educational Measurement, 1970, 7, 63-69.

Newble, D.I. "The evaluation of clinical competence." The Medical Journal of Australia, 1976, 2, 180-83.

Newble, D.I., Elmslie, R.G. & Baxter, A. "A problem-based criterion-referenced examination of clinical competence." Journal of Medical Education, 1978, 53, 720-26.

Newble, D.J., Hoare, J. & Elmslie, R.G. "The validity and reliability of a new examination of the clinical competence of medical students." Medical Education, 1981, 15, 46-52.

Oaks, W.W., Scheinok, P.A. & Husted, F.L. "Objective evaluation of a method of assessing student performance in a clinical clerkship." Journal of Medical Education, 1969, 44, 207-13.

O'Donohue, W.J. & Wergin, J.F. "Evaluation of medical students during a clinical clerkship in internal medicine." Journal of Medical Education, 1978, 53, 55-58.

Printen, K.J., Chappel, W. & Whitney, D.R. "Clinical performance evaluation of junior medical students." Journal of Medical Education, 1973, 48, 343-48.

Sheehan, T.J., Husted, S.D.R., Candee, D., Cook, C.D. & Bargen, M. "Moral judgment as a predictor of clinical performance." Evaluation & the Health Professions, 1980, 3(4), 393-404.

Turner, E.V., Helper, M.M., Kriska, S.D., Singer, S.A. & Ruma, S.J. "Evaluating clinical skills of students in pediatrics." Journal of Medical Education, 1972, 47, 959-65.

Van Wart, A.D. "A problem solving oral examination for family medicine." Journal of Medical Education, 1974, 49, 673-80.

Wile, M.Z. "External examinations for internal evaluation: The National Board Part 1, test as a case." Journal of Medical Education, 1978, 53, 92-97.

Wingard, J.R. & Williamson, J.W. "Grades as predictors of physicians' career performance: An intuitive literature review." Journal of Medical Education, 1973, 48, 311-22.