Simulation in Health Sciences Education

Geoffrey R. Norman
McMaster University
Linda J. Muzzin
McMaster University
Reed G. Williams
Southern Illinois University
David B. Swanson
American Board of Internal Medicine

Simulation has had an increasing role in medical and health science education for over two decades due both to new technology and the movement from content-based curricula toward approaches emphasizing problem solving and clinical reasoning. Various simulation methods have been widely researched as a result and pertinent findings are reviewed here. The review focuses on those techniques involving an encounter between a health provider and client. Excluded are simulations of laboratory experiments, physiological functions, or triage exercises involving the treatment of many patients in simulated emergencies. Five types of simulation methods, oral examinations, live simulated patients, mechanical or electronic simulations, and written and computer-based simulation are reviewed. Each class of simulation will be discussed under the following headings:

(1) Fidelity—To what degree does the simulation method resemble a real life experience?

(2) Reliability—How consistently can performance be assessed with each simulation type?

(3) Validity—Does the simulation method have construct, concurrent or predictive validity?

(4) Learning—Does the simulation result in student learning?

(5) Feasibility—Is the method affordable and feasible to implement?

The questions concern dimensions along which different methods can be compared, however it may not always be possible to find one method which ranks high on all dimensions. Improving a method on one dimension often reduces its effectiveness on others. In the abstract, many simulation methods will have a near "zero sum" and their value depends on the specific educational applications.

Oral Examinations

Oral examinations include a variety of techniques that provoke the examinee to simulate clinical reasoning in response to examiners' questions, either at the bedside or in a direct one-on-one encounter. Skills involved include data gathering, interpretation, differential diagnosis, and management.

Fidelity. Claims about the usefulness of oral exams in judging ability to apply knowledge, to problem solve, to respond to dynamic situations, to demonstrate interpersonal skills and professional attitudes are typically based on unsubstantiated impressions or on reports that the exam seems life like (e.g. Van Wart, 1974), rather than on a more systematic content analysis (Levine & McGuire 1970a, 1970b). The issue of fidelity commonly concerns the extent to which the "reasoning" displayed by candidates in an oral exam reflects how they would actually reason when confronted with a clinical case. Oral exam scores may be based partially on behavior unrelated to clinical competence, such as anxiety level (Pokorney & Frazier, 1966; Waugh & Moyse, 1969), the percentage of words contributed by examinees (Evans, Ingersoll & Smith, 1966), examiner visual impressions (Holloway, Collins & Start, 1968) and examinee self confidence (Wigton, 1980). Whether oral examiners are able to use these cues appropriately remains an issue.

Reliability. Most research has focussed on lack of agreement among examiners (Bull, 1956). But the lack of agreement typically found between totally independent ratings of an examinee in two different situations (Hubbard et al., 1973) might have other explanations. The behavior observed may

be different or the prerequisite skills for the solution of each problem may differ. This interpretation is consistent with the finding that pairs of raters observing the same encounter have reliability coefficients of .75 to .89, while examiners observing different sessions have reliabilities of .25 to .45 (Carter, 1962; Wilson et al., 1969; O'Donohue & Wergin, 1978). While the former may be partially due to consultation between examiners, it does not explain why consensus is often unpredictable among three or more raters (Colton & Peterson, 1967). The problem may result from nonconformists on rating teams whose removal might improve consistency (Newble, Hoare & Sheldrake, 1980; Lloyd, 1983). When checklists and rating forms are used, inter-rater reliability has ranged from .79 to .92 (Maatsch, 1980; Littlefield, Harrington & Garman, 1977) and in general, more rigid structured orals and shorter orals have higher consistency (Bull, 1956; Wilson et al., 1969; Levine & McGuire, 1970b). Longer orals, highlighting diverse skills may reflect "content specificity," or the tendency for mastery of one skill domain to be unrelated to mastery of another, a problem which occurs with all evaluations of competence using real patients or simulation methods.

Validity. Studies of construct validity of medical orals (Miller, 1968; Maatsch, 1980), show differences in performance within levels of training greater than differences between levels. Oral exam validation has tended to rely exclusively on how scores achieved on multiple-choice tests and orals intercorrelate. Although use of written tests as a criterion for validating oral exams is somewhat anomalous, multiple choice tests are highly reliable, widely administered, known to measure factual knowledge and are more familiar than other measures. Most studies show small positive correlations between multiple-choice tests and oral exams (e.g., Bull, 1956; Ludbrook & Marshall, 1971; O'Donohue & Wergin, 1978). These results are often interpreted as evidence that orals and written examinations measure different aspects of clinical competence. However, test unreliability could also reduce these correlations (Levine & McGuire, 1970a; Meskauskas, 1975).

Learning. The oral exam has often been cited as a flexible technique allowing direct feedback between teacher and student and it is institutionalized in most medical schools at the bedside in the form of medical rounds and evaluative clinical exercises in residency programs and clerkships (Futcher, Sanderson & Pusler, 1977). With a few exceptions however, (Vu, Johnson & Mertz, 1981; Powles et al., 1981) the learning value of these exercises has not been reported in the literature.

Feasibility. Oral exams are not feasible for large groups of learners because of logistical problems and costs resulting from having examinees at one location examining the same patients and being rated by the same examiners, now the widely accepted basis for standardization. Both the National Board of Medical Examiners and the American Board of Internal Medicine have had to discontinue orals because of their unreliability and expense.

Discussion. There may be trade-offs between reliability and content validity. As Perimutter remarked, "flexibility is perhaps inversely proportional to standardization" (American Board of Medical Specialties, 1983, p.64). As examinations are made uniform to achieve consistency, the flexibility, uniqueness and spontaneity of the oral exam is lost. High reliability resulting from standardization may only allow sampling a very small part of overall clinical competence, while longer and less structured orals might measure more skills but with lower reliability (O'Donohue & Wergin, 1978). Another issue concerns examiner ability to judge whether an examinee will make a good doctor. Although training and checklists can improve reliability, it is not clear that the judgments are more accurate than judgments that do take place freely or intuitively.

Five simulation methods are reviewed: oral examinations, live simulated patients, mechanical or electronic similations, written, and computer-based similation.

Simulated Patients

Although healthy individuals acting as patients have always had a role in teaching the clinical examination, Barrows (1971) was the first to formally train people to simulate all aspects of a patient problem—history, physical findings and emotions. Simulated patients are purported to have several advantages over actual patients: they can be scheduled and chosen for their appropriateness to learning objectives; they can simulate acute, serious or rare conditions which might not usually be encountered; free discussion can be conducted in their presence that might be viewed as unethical or impolite with real patients; junior students can approach and examine them without concern for physical or emotional harm; and simulated patients can provide accurate and objective feedback without fear of consequences. This approach has been extended to real patients with chronic conditions trained to simulate their manner when they first presented their problem to a health professional (Stillman et al., 1980).

Fidelity. Trained simulated patients have been introduced into family physicians' offices without detection (Burri, McCaughan & Barrows, 1976; Owen & Winkler, 1974). However, other studies have resulted in a detection rate of 20% (Neufeld et al., 1983). Focussing more directly on behavior with simulated and real patients, Norman and Tugwell (1982) found no difference in history-taking, physical examination and diagnosis.

Reliability. The reliability of examinee performance in encounters with simulated patients depends on the scoring method. Subjective ratings using broad categories and five point scales yield inter-rater reliability coefficients in the range 0.5 to 0.7 (Finkel & Norman, 1973). Other approaches have much higher reliability. Because the patient problem is standardized, a criterion group can determine what options on history, physical, and so on constitute an appropriate workup of the patient's problem, and agreement of better than 95 percent (kappa ^{5} .86) between recall of the simulated patient and observer rating on these checklists has been demonstrated (Neufeld et al., 1983).

Other studies report similarly high reliability, with phi coefficients ranging from .77 to .89 (Stillman et al., 1980). However, several studies have shown low correlations of performance on different problems. In one study, the correlation across seven gastro-intestinal problems was only 0.16 (Killer et al., 1983). Another study found correlations ranging from around 0.3 for two presentations of the same problem to 0.1 for problems in different specialties (Norman et al., 1983). Similar results, with correlations from -.26 to ^{s}.22, were found by Rutala et al., (1980). Thus, content specificity once more appears to be a problem common to nearly all evaluations based on a single patient encounter.

Validity. Validity studies have examined how individual performance in simulated patient encounters correlates with performance on other measures. Content-specificity places an upper limit on validity since a stable estimate of performance based on simulated patients requires a large number of encounters. One study (Wakefield & Norman, 1977) compared performance of family medicine residents in an average of twenty-two observed encounters with their supervisor performance reports and their later performance on the certification examination. Scores on data gathering were uncorrelated with either criterion. Clinical judgment correlated .51 with the same competency assessed by supervisors, and 0.54 with examiners ratings of treatment plans in the certification oral. Interpersonal and interview skills correlated 0.3 to 0.4 with simlar categories assessed by supervisor report and oral examinations. Positive correlations have been found between scores of content assessed by patients trained as simulators and peer ratings in data acquisition, problem recognition, attitudes, and competence, supporting concurrent validity. But two studies (Killer et al., 1983; Norman et al., 1983 show no significant correlation between performance on simulated patients and multiple choice test scores. Scores from simulated patient encounters might primarily reflect interpersonal skill unrelated to content knowledge.

Learning. In contrast to reliability and validity, numerous studies show simulated patients have positive educational effects. Several studies have contrasted learning from simulated and real patients (Tinning, 1975; Livingstone & Ostrow, 1978; Holzman et al., 1977; Johnson, Murchison & Reiter, 1976) and students taught with simulated patients have consistently out-performed those taught with real patients.

Feasibility. A major barrier to feasibility is the cost of training and using simulated patients, which requires use of faculty resources. Training and application costs are consistently overestimated, however. At McMaster University, simulated patients are trained in two to three hours at a cost of $10 per hour, and cost of application is also $10 per hour, which compares favorably with that of computer simulations. The real cost is faculty time required to observe and evaluate simulated patient encounters, but there has been significant progress towards using simulated patients without the requirement for faculty participation.

Discussion. It is clear from this review that simulated patients have a strong, positive effect on learning. Further, the approach has demonstrated validity in comparison to real patients which is unequaled by other simulation methods. Both these features suggest a useful role for live simulated patients in teaching and evaluation which is not, as yet, fully realized.

All methods of evaluation must deal with the problem of "content specificity," the tendency for mastery of one skill domain to be unrelated to mastery of another.

Manikins

Perhaps the most widespread but least discussed simulation devices in health sciences education are manikins such as RESUSCI-ANNIE (for cardiopulmonary resuscitation) and GINNIE (for pelvic examination). The term "manikin" is an over-simplification. Although some are the essence of simplicity, others, like the SIM-1 computer-controlled manikin (Abrahamson, Denson & Wolf, 1969) or the cardiac simulator HARVEY (Gordon, 1981), are quite complex. Most are mechanical or electronic, but also included are animal "models" such as anesthetized cats for teaching pediatrics intubation (Thompson, 1979). Manikins are designed universally to teach and evaluate a defined procedural or technical skill. In most applications, the focus is on learning rather than assessment. Some controlled experiments have been conducted with manikins. Evaluation using simulators can be process oriented, by using detailed checklists, or outcome oriented, by comparing the student's formulation (e.g. split third heart sound, retinal exudates) with a known abnormality. Either or both orientations may be appropriate.

Fidelity. Fidelity depends on the specific device. SIM-1, is highly realistic, even to blinking eyes, but other manikins are less realistic. Little empirical evidence of fidelity exists. Thompson (1979) stated that intubation of kittens was highly realistic, which only proves that beauty is in the eye of the beholder, or throat of the beholden. Seventy-five to 85% of physicians found HARVEY realistic (Gordon, 1981). However, fidelity of simulations need only be consistent with the educational purpose (Simon, 1981). For example, airplane simulators need a complete array of instruments and authentic responses to pilot actions but do not need wings to serve their purpose.

Reliability. Performance is usually assessed with detailed behavioral checklists and reliability is usually high. Inter-rater agreements of 95% have been reported with the cardiac simulator HARVEY (Gordon, 1981) while other studies using checklists, without manikins, report reliabilities in the range of .75 to .94 (Liu et al., 1980; Andrews, 1977). Very high reliability is achievable through the use of manikins and behavioral checklists.

Validity. Little empirical evidence of validity exists. Abrahamson et al., (1969) showed that residents trained on SIM-1 performed better than others in the clinical setting, implying some concurrent validity. However a study of a less life-like simulator found that a group trained on a heart sound simulator (Sajid, Magero & Feinzimer, 1977) showed large gains in recognizing taped heart sounds, but there was no dif-

ference in performance with actual patients. Clearly, further research is required.

Learning. Thompson (1979) indicated that residents found practicing intubations on cats "useful". Penta and Kofman (1973) conducted a randomized trial, and showed significant differences between groups using the Iowa Opthalmoscope model and the Strabismus Cover Test Simulator, but no differences for the Bartner eye model or a heart sound simulator. As already noted, students trained on SIM-1 and on a heart sound simulator showed superior performance, but only SIM-1 training transferred to clinical practice.

Feasibility. Widespread use of simple models like RESUSCI-ANNIE and heart sound simulators attest to their feasibility. Conversely, the enormous development costs of computer controlled manikins like SIM-1 has undoubtedly led to few adoptions. Given their variety, a global judgment about the feasibility of manikins cannot be made.

Discussion. Manikins, or technical simulators have achieved a limited, but important, role in teaching clinical skills. When defined technical or motor skills are to be acquired, these simulation devices can be useful. However, it is discouraging to see limited rigorous evaluation of their validity and effectiveness.

Written and Computer-Based Simulations

Written and computer/based simulations have been used to assess problem solving since the late 1960s and have been investigated in many studies. Findings from research on both kinds of simulation are deliberately integrated since most work with computer simulation has simply replaced paper and pencil with a computer terminal, with little evidence indicating that computer use has important measurement consequences. Recent applications of computer simulation, using microcomputer and videodisc technology does represent significant progress but as yet no studies of these methods have been reported.

Fidelity. Several studies have investigated whether actions taken on these simulations are similar to those which occur with real patients. Three different research paradigms have been used—introspection/self report, chart audit, and live simulated patients. Introspection/self report studies (e.g. McGuire & Babbott, 1967, Harless et al., 1978; Finchman et al., 1976; National Board of Medical Examiners and American Board of Internal Medicine, 1981) indicate that examinees consider the simulations realistic. Audit studies, comparing behavior on the simulations with similar real life cases, show performance similarity, at least in selection of important diagnostic and therapeutic actions (Goran, Williamson & Gonnella, 1973; Harless et al., 1978; National Board of Medical Examiners and American Board of Internal Medicine, 1981), but more data is collected using formats that cue the examinee by delineating specific options. Comparisons of behavior on written and computer-based simulations with identical cases using simulated patients have also demonstrated cueing effects (Feightner & Norman, 1978; Norman & Feightner, 1981; Page & Fielding, 1980), which lead to higher scores than in uncued formats or real life. Cueing problems may be worse in the written than computer formats because examinees may read ahead and use choices listed in later sections to determine which earlier options are appropriate.

Reliability. Three sources of measurement error have been investigated—variation introduced by the scoring procedure used, inter-rater variation in developing scoring keys and intercase variation in performance. Most scoring schemes for written and computer simulations involve categorizing each option (e.g. essential, indicated, contraindicated and risky) and assigning a numerical weight to each category (e.g. essential =+2, indicated =+1, contraindicated =+1, and risky =+2). Changes in number of categories and moderate changes in the weights assigned categories make little difference in resulting scores (Bligh, 1980, Norcini et al., 1983a). Correlations between such score variants are so high that reliability and validity are unaffected. These results mirror research on weighted versus unweighted composite scores (Lord & Novick, 1968) and prediction accuracy using linear composites with different regression weights (Dawes & Corrigan, 1974).

Poor inter-rater agreement on appropriate (and inappropriate) options adversely affects scoring clinical simulations. There can be substantial variation in workup and treatment recommendations among skilled physicians on real and simulated cases (Barrows et al., 1978; Elstein, Shulman & Sprafka, 1978). As a result, consensus judgment by an expert panel seems necessary and appropriate for constructing scoring keys (Mazzuca & Cohen, 1982), but it is unclear if consensus scoring adequately rewards or penalizes approaches to patient management that reflecting differences in quality of care versus differences in style and opinion.

Previous studies have shown that clinical problem solving is not a general trait but is specific to particular content domains (e.g. diabetes, coma, emotional problems). Physicians have abilities which vary from specialty, and indeed, from case to case within a specialty (Elstein, Shulman & Sprafka, 1978; Barrows et al., 1978; Swanson et al., 1982). This is reflected psychometricly in low intercorrelations between scores on different cases and poor reliability of composite test scores. Most correlations are from 0.1 to 0.4 (Marshall, 1977; Berner, Bligh & Guerin, 1977; Mast et al., 1982) indicating that performance on one case is a poor predictor of performance on other cases. Therefore, a large number of cases are needed for acceptable levels of intercase reliability. If intercase correlations average 0.2, and a total test reliability of 0.8 is desired (a minimum value for assessment of individual performance), then about 16 cases are required. Most tests using written or computer based simulations are significantly shorter than is psychometrically reasonable, and yield scores with marginal reliability.

Validity. Construct validity studies have examined the relationship between different subscores on the simulation using medical problem solving theories to guide interpretation (Juul, Noe & Nereberg, 1979; Berner, Bligh & Guerin, 1977; Harasym et al., 1980). Results

> The use of simulated patients demonstrates a validity which is unequalled by other simulation methods.

have varied since each uses somewhat different subjects, measures, and analytic procedures, but most use factor analysis as a "discovery procedure" to identify underlying dimensions of problem solving performance. Investigators have found 1) a single, weak underlying factor, 2) two underlying factors typically labeled data-gathering and decision-making, or 3) no interpretable factor. Exploratory factor analysis may be inappropriate, given usually small sample sizes and low intercase correlations which affect factor stability.

Correlations between written or computer simulation performance and other written clinical competence measures, particularly multiple choice test scores, are low to moderate ranging from 0.2 to 0.5 (Case, 1981; Langdon et al., 1978; McGuire & Babbott, 1967). As with oral exams, this is interpreted to mean that simulations measure something different, but given the low intercase reliability of the typical simulation-based exam, low intercorrelations may simply reflect attenuation due to measurement error. A study of the day-long certifying exam in internal medicine using three years of data from five thousand subjects per year,(Norcini, Swanson & Webster, 1983) found correlations ranging from 0.71 to 0.76 between the multiple-choice and simulation components and true correlations greater than 0.9, strongly supporting the attenuation interpretation. Thus, no conclusions regarding construct validity can be drawn because of wide variations in results and attenuation of correlations. More research and better analytic procedures are badly needed.

Criterion-referenced validity has been investigated using the relative performance of groups which should differ in ability and correlations with some performance-based measure of ability, usually ratings from clinical instructors. Written simulations can distinguish performance of medical students and residents at different training stages (McGuire & Babbott, 1967; Schumacher, 1974; Schumacher et al., 1974; National Board of Medical Examiners and American Board of Internal Medicine, 1981; Grosso & Webster, 1983). Given that residents have a year or two of additional education and clinical experience, this is not surprising. A well constructed twenty-item multiple choice test could probably demonstrate similar differences so only mild support for validity is implied by these results.

Studies relating simulation scores to real-world criterion measures of clinical ability such as clinical instructor ratings have found correlations in the 0.1 to 0.4 range (Langdon et al., 1978; Norcini, Swanson & Webster, 1983; Schumacher, Burg & Taylor, 1974), approximating correlations between multiple-choice tests and similar real world criterion measures. Studies of incremental validity— whether predictions of the criterion improve when written simulations are added to a multiple-choice test battery—typically show modest improvements (Schumacher, 1974; Langdon et al., 1979; Norcini, Swanson & Webster, 1983b), suggesting that multiple choice tests and simulations measure similar competencies. Again, the low reliability of simulation-based exams make these results difficult to interpret and lack of a good criterion measure of real world clinical performance obscures interpretation further, particularly since instructor ratings are typically plagued by reliability problems.

Manikins have an important role in teaching clinical skills, but there has been only limited rigorous evaluation of their validity and effectiveness.

It is often assumed that written and computer based simulations possess a high degree of content validity because of their high fidelity. Although the evidence is not overwhelming a number of investigators have asserted that the similarity of the problem solving tasks in patient care and written simulations is sufficiently great that the simulations are content valid (McGuire & Babbott, 1967; Swanson et al., 1982). Evidence of consistency of behavior on similar real or simulated cases is, indeed, relevant to an assessment of content validity, but equating content validity with fidelity is incorrect. For a test to be content valid, it is important that the simulated cases be sampled from the real world clinical domain in such a way that they are representative. Judgment enters into the choice and poor case selections can easily be made. Also, the discouraging intercase reliability findings reviewed previously indicate that a large sample of cases is necessary to develop an accurate ability estimate. Tests using small numbers of cases cannot be content valid, because they do not sample clinical situations sufficiently.

Feasibility. The widespread use of written simulations by specialty boards, licensing bodies, and medical schools attests to their feasibility. They are fairly inexpensive to design, can be administered to large numbers of examinees and machine-scored at relatively low cost. Conversely, computer simulations have not proceeded beyond pilot testing or small scale implementation despite years of development. Estimates of development costs range from a few thousand to $50,000 per problem, and costs of administration have been estimated at $50 to $1,000 per examinee (Diamond & Weiner, 1974; Senior, 1976).

It must, however, be recognized that two factors may alter these figures. First, computer equipment costs have decreased dramatically since the mid-1970s and are not reflected in these estimates, and second, computer simulation, perhaps more than any other simulation methodology, is amenable to an economy of scale so that future widespread dissemination may be accompanied by significant cost reductions per examinee.

Learning. Some authors have suggested that written and computer simulations may have a useful role in stimulating learning, however, in this review no studies were identified demonstrating learning gain as a result of these experiences.

Discussion. Despite their widespread use, written and computer simulations have two serious shortcomings—content specificity, which may be a generic problem of all evaluation based on a single patient encounter, and fidelity, since large differences exist between how people perform on written simulations and with actual patients or live simulations. These deficiencies could be overlooked if it could be convincingly demonstrated that the simulations

measure a distinct and central component of competence; however, such evidence is lacking. The value of computer simulations is more problematic, since early developments apparently possessed all of the problems of written simulations and the additional handicap of high development costs, but new developments such as videodisc technology may represent a real advantage in evaluation which remains to be demonstrated.

Conclusions

Content specificity seems to be a fundamental problem in assessing clinical problem solving ability with simulations. Even with very high fidelity computer simulations, it can be anticipated that correlations between performance on different cases will be low. This seems to be a characteristic of problem solving in real clinical life, so it can be expected on simulations as well. It is necessary to use large numbers of cases to adequately assess problem solving ability. Measuring performance on a single case with more fidelity and accuracy can well result in a less valid test because more testing time is usually required. Certainly, short simulation-based exams, consisting of only a few cases, should not be used for either testing or research purposes. One line of inquiry which may reduce the demands for a large number of cases in testing situations is directed at a better understanding of the characteristics of physician knowledge about medical problems. Bordage & Allen (1982) have made an important step in this direction, and other recent papers address the issue (Berner, 1984).

Despite shortcomings, simulations have a significant role to play in health sciences curricula. Live simulations are a powerful method to teach data gathering, interpersonal skills and technical skills, and can be usefully employed to evaluate these skills provided sufficient numbers of patient problems are employed. Some types of manikins have been shown to be useful for both teaching and evaluation specific technical or procedural skills, although further research is necessary to determine those characteristics of the simulation which result in effective transfer to the clinical setting. Finally, computer simulations have had an uncertain role in the past, but the evolution of new technology holds great promise for the future.

No studies have been reported of medical simulations that use recent advances in microcomputer and videodisc technology.

References

Abrahamson, S., Denson, J.S., & Wolf, R.M. (1969). Effectiveness of a simulator in training anesthesia residents. *Journal of Medical Education, 44,* 515-518.

American Board of Medical Specialties. (1983). *Oral Examinations in Medical Specialty Board Certification.* J.S. Lloyd, (Ed.). Chicago, Il: Author.

Andrews, B.J. (1977). The use of behavioral checklists to assess physical examination skills. *Journal of Medical Education, 52.* 589-590.

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

Berner, E.S., Bligh, T.J., & Guerin, R.O. (1977). An indication for a process dimension in medical problem-solving. *Medical Education, 11,* 324-328.

Berner, E.S. (1984). Paradigms and problem solving: A literature review. *Journal Medical Education, 59,* 625-633.

Bligh, T.J (1980). *Written simulation scoring: A comparison of nine systems.* Presented at the annual meeting of the American Educational Research Association, Boston.

Bordage, G., & Allen, T. (1982). The etiology of diagnostic errors: Process or content? An exploratory study. *Proceedings of the 21st conference on research in medical education,* Washington, DC.

Bull, G.M. (1956). An examination of the final examination in medicine. *Lancet, 271,* 368-372.

Burri, A., McCaughan, & Barrows, H. (1976). Using simulated patients to evaluate practicing physicians in a community. *Proceedings of the 15th conference on research in medical education,* San Francisco, CA.

Carter, H.D. (1962). How reliable are good oral examinations? *California Journal of Education Research, 13,* 147-153.

Case, S.M. (1981). A new examination for the evaluation of diagnostic problem solving. *Proceedings of the 20th conference on research in medical education,* Washington, DC.

Colton, T., & Peterson, O.L. (1967). An essay of medical students' abilities by oral examination. *Journal of Medical Education, 42,* 1005-1014.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychology Bulletin, 81,* 95-106.

Diamond, H.S., & Weiner, M. (1974). A computer assisted instructional course in diagnosis and treatment diseases. *Arthritis and Rheumatism, 17,* 1049-1054.

Elstein, A., Shulman, L.S., & Sprafka, S.A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Evans, L., Ingersoll, R.W., & Smith, E.J. (1966). The reliability, validity and taxonomic structure of the oral examination. *Journal of Medical Education, 41,* 651.

Feightner, J.W., & Norman, G.R. (1978). Computer-based problems as a measure of the problem-solving process: Some concerns about validity. *Proceedings of the 17th conference on research in medical education,* Washington, DC.

Finchman, S.M., Grace, M., Taylor, W.C., Dsadun, E.N., & Davis, F.C. (1976). Pediatric candidates' attitudes to computerized patient management problems in a certifying examination. *Medical Education, 10,* 404-407.

Finkle, A., & Norman, G.R. (1973). The validity of direct observation. *Proceedings of the 12th conference of research in medical education,* Washington, DC.

Futcher, P.H., Sanderson, E.V., & Pusler, P.A. (1977). Evaluation of clinical skills for a specialty board during resident training. *Journal of Medical Education, 52,* 567-577.

Goran, M.J., Williamson, J.W., & Gonnella, J.S. (1973). The validity of patient management problems. *Journal of Medical Education, 48,* 171-178.

Gordon, M. (1981). HARVEY: The cardiology simulator. *Journal of Practice, 13,* 353-357.

Harasym, P., Baumber, J., Fundytus, D., Preshaw, R., Watanbe, M., & Wyse, G. (1980). An evaluation of the clinical problem-solving process using a simulation technique. *Medical Education,14,* 381-386.

Harless, W.G., Farr, N.A., Zier, M.A., & Gamble, J.R. (1978). A method for physician recertification. *T.H.E. Journal, 5,* 51-55.

Holloway, P.J., Collins, D.K., & Start, K.B. (1968). Reliability of viva voce examinations. *British Dental Journal, 125,* 211-214.

Holzman, G.B., Singleton, D., Holmes, T.F., & Maatsch, J.L. (1977). Initial pelvic examination instruction: The effectiveness of three contemporary approaches. *American Journal of Obstetrics and Gynecology, 6,* 129, *5,* 124-129.

Hubbard, J.P, Levitt, E.J, Schumacher, C.F., & Schnable, T.G. (1973). An objective evaluation of clinical competence. *New England Journal of Medicine, 272,* 1321-1328.

Johnson, C.F., Murchinson, N., & Reiter, S. (1976). Sick infants versus simulated well-baby examination as an initial pediatric learning experience. *Journal of Medical Education, 6(51), 1021-1023.*

Juul, D.H., Noe, M.J., & Nerenberg, R.L. (1979). A factor analytic study of branching patient managemant problems. Medical Education, 13(3), 199-203.

Killer, D., Tugwell, P., & Norman, G.R. (1983). Stability of scores in resident performance on gastrointestinal problems. Unpublished manuscript, McMaster University.

Langdon, L.O., Maskauskas, J.A., Norcini, J.J., & Webster, G.D. *ABIN MERIT: Recertification na-*

tional study. American Board of Internal Medicine, Internal Report.

Levine, H.G., & McGuire, C.H. (1970). The use of role-playing to evaluate effective skills in medicine. *Journal of Medical Education, 45,* 700-705.

Littlefield, J.H., Harrington, J.T., & Garman, R.E. (1977). Use of an oral examination in an internal medicine clerkship. *Proceedings of the 16th conference on research in medical education,* Washington DC.

Liu, P., Miller, E., Herr, G., Hardy, C., Sivarajan, M., & Willenkin, R. (1980). Videotape reliability: A method of evaluation of a clinical performance examination. *Journal of Medical Education, 55,* 713-715.

Livingstone, R., & Ostrow, D.N. (1983). Professional patient instructors in the teaching of the pelvic examination. *American Journal of Obstetrics and Gynecology,* 132, 54-57.

Lloyd, J.S. (Ed.). (1983). *Oral examination in medical specialty board certification.* Chicago, American Board of Medical Specialties.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing.

Ludbrook, J.J., & Marsholl, V.R. (1971). Examiner training for clinical examinations. *British Journal of Medical Education, 6,* 152-155.

Maatsch, J.L. (1980). *Model for criterion-referenced medical specialty test.* Office of Medical Education research and Development Michigan State University in collaboration with the American Board of Emergency Medicine.

Marshall, J. (1977). Assessment of problem-solving ability. *Medical Education, 11,* 329-334.

Meskauskas, M.S. (1975). A study of the oral examinations of the Subspecialty Board of Cardiovascular Disease of the American Board of Internal Medicine. *Proceedings of the conference on oral examination,* Des Plaines, IL: American Board of Medical Specialties.

Mast, T.A., Colliver, J.A., Anderson, M.B., & Soler, N.G. (1982). Validation of problem solving measures: Multitrait-multimethod matrix analysis. *Proceedings of the 21st conference on research in medical education,* Washington, DC.

Mazzuca, S.A., & Cohen, S.J. (1982). Scoring patient management problems, external validation of expert consensus. *Evaluation and Health Professions, 5,* 210-217.

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

Miller, G.E. (1968). The orthopaedic training study. *Journal of the American Medical Association, 206,* 601-606.

Neufeld, V.R., Woodward, C.A., & Norman, G.R. (1983). Simulated patients in evaluation of medical education. *Proceedings 22nd conference on research in medical education,* Washington, DC.

Newble, D.I., Hoare, J., & Sheldrake, P.F. (1980). The selection and training of examiners for clinical examinations. *Medical Education, 14,* 345-349.

Norcini, J.J., Swanson, D.B., Grosso, L.J., & Webster, G.D. (1983). A comparison of several methods for scoring patient management problems. *Proceedings of the 22nd conference on research in medical education,* Washington, DC.

Norcini, J.J., Swanson, D.B., & Webster, G.D. (1983). Reliability validity, and efficiency of various item formats in assessment of physician competence. *Proceedings of the 22nd conference on research in medical education,* Washington, DC.

Norman, G.R., Feightner, J.W., Tugwell, P., Muzzin, L.J., & Guyatt, G. (1983). The generalizability of measures of clinical problem-solving. *Proceedings of the 22nd conference on research in medical education,* Washington, DC.

Norman, G.R., & Feightner, J.W. (1981). A comparison of behavior on simulated patients and patient management problems. *Medical Education, 15,* 26-32.

Norman, G.R., & Tugwell, O. (1982). A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 57,* 708-715.

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

Owen, A., & Winkler, R. (1974). General practitioners and psychosocial problems: An evaluation using pseudopatients. *Medical Journal of Australia, 2,* 393-398.

Page, G.G., & Fielding, D.W. (1980). Performance on PMPs and performance in practice: Are they related? *Journal of Medical Education, 55,* 529-537.

Penta, F.B., & Kofman, S. (19730* The effectiveness of simulation devices in teaching selected skills of physical diagnosis. *Journal of Medical Education, 43,* 442-445.

Pokorney, A.D., & Frazier, S.H (1966). An evaluation on oral examinations. *Journal of Medical Education, 41,* 28-40.

Powles, A.C.P., Wintrip, N., Neufeld, V.R., Wakefield, J.G., Coates, G., & Burrows, J. (1981). The "triple-jump" exercise: Further studies on an evaluative technique. *Proceedings of the 20th conference on research in medical education,* Washington, DC.

Rutala, P.J., Stillman, P.L., & Sabers, D.L. (1980). Patient instructors as evaluators of housestaff clinical competence. *Proceedings of the 19th conference on research in medical education,* Washington, DC.

Sajid, A., Magero, J., & Feinzimer, M. (1977). Learning effectiveness of the heart sound simulator. *Journal of Medical Education, 11,* 25-27.

Schumacher, C.F. (1974). A comparative study of four methods for scoring on experimental computer-based examination for clinical problem solving. *Proceedings of the 13th conference on research in medical education,* Washington, DC.

Schumacher, C.F., Burg, F.D., & Taylor, W.C. (1974). computerization of a patient management problems examination to prevent retracing. *Proceedings of the 13th conference on research in medical education,* Washington, DC.

Senior, J.R (1976). *Toward the measurement of competence in medicine.* Philadelphia: American Board of Internal Medicine.

Simon, H.A. (1981). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Stillman, P.L., Ruggil, J.S., Rutala, P.J., & Sabers, D.L. (1980). Patient instructors as teachers and evaluators. *Journal of Medical Education, 55,* 186-193.

Swanson, D.B., Barrows, H.S., Friedman, C.P., Levine, H.G., & Norman, G.R., (1982). Issues in assessment of clinical competence. *Professions Education researcher Notes, 4.*

Thompson, B. (1979). Use of kittens in teaching neonatal resuscitation to family medicine residents. *Journal of Family Practice,9,* 128-129.

Tinning, F.C. (1975). *Simulation in Medical Education,* Unpublished doctoral thesis. Michigan State University.

Van Wart, A.D. (1974). A problem-solving oral examination for Family Medicine. *Journal of Medical Education, 49,* 673-679.

Vu, N.V., Johnson, R., & Mertz, S.A (1981). Oral examination: A model for its use within a clinical clerkship. *Journal of Medical Education, 56,* 665-667.

Wakefield, J.G., & Norman, G.R. (1977). Assessment of cognitive and interpersonal skills in clinical problem solving. *Proceedings of the 16th conference of research in medical education,* Washington, DC.

Waugh, D., & Moyse, C.A. (1969). Medical education II: Oral examinations: A video study of the reproducibility of grades in pathology. *Canadian Medical Association Journal, 100,* 635-640.

Wigton, R.S. (1980). The effects of student personal characteristics on the evaluation of clinical performance. *Journal of Medical Education, 55,* 423-427.

Wilson, G.M., Harden, R. McG, Lever, R., Robertson, J.I.S., & MacRitchie, J. (1969). Examination of clinical examiners. *Lancet, 296,* 37-40.