A PRELIMINARY STUDY OF THE IMPACT OF CASE SPECIFICITY ON COMPUTER-BASED ASSESSMENT OF MEDICAL STUDENT CLINICAL PERFORMANCE

JAMES T. FITZGERALD
FREDRIC M. WOLF
WAYNE K. DAVIS
MEL L. BARCLAY
MARY E. BOZYNSKI
KENNETH R. CHAMBERLAIN
University of Michigan Medical School
STEPHEN G. CLYMAN
National Board of Medical Examiners
THOMAS C. SHOPE
JAMES O. WOOLLISCROFT
GERALD B. ZELENOCK
University of Michigan Medical School

The implications of case specificity of two computer-based clinical simulation examination cases (CBX) were examined by a classical measurement approach and by a Bayesian analysis of test characteristics. The CBXs (a surgery and an ob/gyn case) were designed by the National Board of Medical Examiners and administered to 163 University of Michigan Medical School students. The results indicate that the students performed differently on the two cases, the surgery case appearing to be more difficult. The ob/gyn case had greater sensitivity (more accuracy in passing competent students), whereas the surgery case had greater specificity (more accuracy in failing noncompetent students). The differences between the cases and evidence of case specificity raise the issue of an exam's objective and the acceptable type of classification error. These results suggest that additional studies are required before widespread use of such exams can be implemented in "high stakes" situations for licensure purposes.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 17 No. 3, September 1994 307-321

Evidence of case specificity in problem-solving performance has raised a number of issues and potential difficulties in the assessment of clinical performance, as well as in design of learning activities and curriculum. These issues include the reliability of scores (particularly for high stakes examinations) (Swanson & Stillman, 1990), the minimum number of cases needed to assess competence reliably and validity (van der Vleuten & Swanson, 1990), the nature of cognitive activities manifested (or latent) in problem solving and clinical reasoning, and the degree of transferability from one task to another (Elstein, Shulman, & Sprafka, 1978, 1990). Because the performance of an individual across clinical problems has been shown to be inconsistent, for example, an expert in one medical domain is not likely to be an expert in other medical domains, it is important to understand the implications of case specificity for curriculum and assessment purposes. The present study examined medical students' clinical performance on two different computer-based simulations and discusses the implications of differential performance for assessment purposes. Classical measurement approaches were used in addition to a Bayesian analysis of test characteristics, including test sensitivity, test specificity, and true and false positive rates. The two approaches are complementary, each providing unique information and understanding on the implications of performance on simulated cases.

The National Board of Medical Examiners (NBME) is currently developing computer-based clinical simulation examinations (CBX) to assess clinical skill performance for use in licensing exams. A CBX case introduces a patient problem that an examinee must diagnose and manage over time. It has been suggested that CBX problems measure aspects of clinical performance that are different from the skills and knowledge tested in more traditional pencil and paper exams, for example, the assessment of clinical judgment (Volle, 1990). One advantage is that in a CBX case the management and treatment of a patient can easily be introduced in a no risk situation. Simulations can also have a clock component that allows the time to advance after decisions or actions of the examinee so that performance in management can be ascertained.

In 1991, the University of Michigan Medical School (UMMS) conducted a Comprehensive Clinical Assessment (CCA) of medical

students at the end of their third year of training. During the 2 weeks preceding the CCA, students were required to complete two CBX cases. It was hoped that performance on the two CBX cases would provide an evaluation of students' problem-solving and patient management abilities. The CBX was limited to two cases due to time and resource constraints. Approximately 2 months after the CBX, the NBME Part 2 exam was taken by the students.

In this study, the predictive validity and the case specificity of the two cases were compared. Student performance and case differences were evaluated by two methods. The first method assessed the ability of the two CBX scores to predict scores in subsequent assessments of clinical skills/knowledge (the CCA and the NBME Part 2). The predictive validity of the CBX cases was also compared to measures more distant in time (e.g., MCAT average) and measures lacking a direct focus on clinical skills (e.g., number of unacceptable grades during medical school).

The second method for examining case specificity used a Bayesian analysis typically used to evaluate clinical test characteristics. The sensitivity and specificity of both CBX cases were assessed first using CCA performance and then NBME Part 2 performance as the "gold standard" measures of competency. Comparing student performance on each case to external classifications of competency allowed us to determine the similarities or differences in the assessment capabilities of the two CBX cases.

Combining CBX scores was considered and analyses performed using a combined score. However, because of differences in the scoring scales and the degree of difficulty between the two cases, we felt the results of any analysis using a combined score would be difficult to interpret, may mask important characteristics of the individual cases, and would possibly be misleading. Furthermore, it is hard to defend a total score using only two cases. Rather, we felt it was more important to look at student performance on the two individual cases in order to better understand the implications for competency assessment purposes.

Using these methods to determine the predictive validity and the case specificity of the two CBX cases, the following research questions were explored:

- What are the best predictors of CCA performance among the measures of CBX scores, undergraduate GPA, MCAT score, NBME Part 1 score, and the number of unacceptable grades during medical school?
- What are the best predictors of NBME Part 2 performance among the measures of CBX scores, undergraduate GPA, MCAT score, NBME Part 1 score, and the number of unacceptable grades during medical school?
- Are there differences in the two CBX cases in the prediction of the CCA and NBME Part 2 performance?
- Are there differences in the ability of the two CBX cases to correctly identify competent students (sensitivity) or to correctly identify non-competent students (specificity)?

METHOD

STUDY PARTICIPANTS

This study examined two CBX cases designed by the NBME and administered to 163 standard program medical students at UMMS. Of these students, 28% were women and 22% were minorities underrepresented in medicine. In July 1991, students signed up for one of 24 testing sessions. These sessions took place over 7 days during a 2 week period. A maximum of 10 students were tested individually in each session. The students were given an hour to complete each case. Given the time constraints and the number of students that needed to complete the CBX cases, only two different cases were assigned to each student.

A CBX orientation module was available to all students prior to the testing period. Although many students had been introduced to CBX during their ob/gyn clerkship, students were encouraged to complete the orientation program. Thus the vast majority of the students had exposure to CBX prior to the their testing session. Staff were also available to assist the students if computer problems arose during the test. Therefore, we believe that score variance due to unfamiliarity with CBX was kept to a minimum and had little impact on the study's results.

COMPUTER-BASED CLINICAL SIMULATION EXAMINATION (CBX)

The history and development of the NBME's CBX is provided by Clyman & Orr (1990). The purpose of the CBX is to evaluate clinical competence in a realistic, unprompted setting through a patient problem that evolves over time. A CBX case simulates a patient and physician interaction, beginning with a short vignette that introduces a patient and the chief complaint. The student then makes requests to obtain patient information, orders tests, and begins treatment of the patient. The student is expected to diagnose and manage the patient's health problem over time. In this assessment, students were assigned a surgery and an ob/gyn CBX.

A student's information requests, management decisions, and progress through the simulation are charted and scored. For each case, two evaluations are provided: an overall score and the number of flags received. The overall score is an evaluation of the student's performance on the case as a whole. The higher the overall case score, the better the student performed. Flags represent inappropriate actions or omissions and indicate a serious misunderstanding of the case. Because flag actions or omissions may be dangerous for the patient, it is desirable to complete a case without receiving a flag.

COMPREHENSIVE CLINICAL EXAMINATION (CCA)

The CCA is a multiple station clinical examination given to the UMMS students at the end of third year clerkships. The exam is designed to determine if a student has the knowledge and skills expected of students entering their fourth year of medical school. The exam tests clinical skills, such as history taking, x-ray interpretation, EKG interpretation, and so on. The exam is composed of 10 stations with a total of 85 questions (an 11th station was not used in the student assessments). The reliability coefficient alpha of the 85 items of the CCA was 0.77. We believe this indicates moderate reliability. The lack of a higher reliability is probably due to methods variance within the exam. The exam used many formats to test performance on a number of clinical tasks. The use of multiple formats had the effect of

suppressing the magnitude of the correlation of the CBX cases with the CCA.

STATISTICAL METHODS

Pearson correlation coefficients were used to examine the relationships among the measures included in the study. To explore the predictive ability/validity of the CBX scores, two regression models were constructed. In the first model, the CCA score was the dependent variable, while the independent variables included undergraduate GPA, MCAT average, the number of unacceptable grades during medical school, the NBME Part 1 score, the overall ob/gyn CBX score, and the overall surgery CBX score. In the second model, the NBME Part 2 score replaced the CCA score as the dependent variable.

The number of flags received for each CBX case was not used in either model because they represent different evaluations of the same exam performance. The overall scores were selected over the flag totals because these scores have a greater range. Also provided in the regression analysis are the standardized regression coefficients (standardized βs) for the independent variables; "these coefficients are independent of the scales of measurement of the independent variables and may offer a comparison of the magnitude of the effects of the variables" (Freund, Littell, & Spector, 1986, p. 26) Standardized βs make possible direct comparisons of the independent variables.

The sensitivity and specificity of the two CBX cases were calculated using flag totals. Sensitivity is a measure of the exam's accuracy in passing individuals who are assumed to be competent. Specificity is a measure of the exam's accuracy in failing individuals assumed not to be competent. To construct the 2×2 tables necessary for this type of analyses, the flag score for each case was recoded as a dichotomous variable. Students receiving zero flags for a case were considered as having passed the case, whereas students receiving one or more flags were considered to have failed the case. The recoded flag scores were cross-tabulated with the CCA, the NBME Part 1, and the NBME Part 2 pass/fail (competent/noncompetent) determinations.

RESULTS

CORRELATIONS

Table 1 presents the correlations among the CBX measures (the overall and the flag score of each case), the CCA, NBME Parts 1 and 2, undergraduate GPA, MCAT average, and the number of unacceptable grades during medical school. The correlation analysis reveals little or no association between the ob/gyn CBX case scores and the non-CBX measures. The surgery scores, however, have significant correlations with all of the non-CBX measures. The strength of the correlations between the overall surgery score and the non-CBX measures range from a high of 0.44 (NBME Part 2) to a low of 0.18 (undergraduate GPA). The correlations of the surgery flag total are similar, ranging from 0.44 (NBME Part 2) to 0.17 (undergraduate GPA). The surgery measures' strongest correlations were with the more medically related measures (the CCA, the NBME Part 1, and the NBME Part 2 scores).

REGRESSION MODELS

In the first regression model, two of the six independent variables were significantly related to the dependent variable, the CCA score (see Table 2). These variables were the NBME Part 1 score and the CBX surgery score. The standardized βs indicated that the NBME Part 1 score, $\beta = 0.45$, was twice as important as the CBX surgery score, $\beta = 0.22$, in the prediction of CCA performance.

The second regression model used the NBME Part 2 score as the dependent variable (see Table 2). Again, the NBME Part 1 and the CBX surgery scores prove to be the only two variables that enter into the prediction of the dependent variable. The standardized βs show a much larger disparity between the NBME Part 1 score, $\beta = 0.74$, and the CBX surgery score, $\beta = 0.11$.

SENSITIVITY AND SPECIFICITY

The results of the sensitivity and specificity analyses are found in Tables 3 and 4. In Table 3, sensitivity and specificity were calculated

TABLE 1

Correlations Among the Measures of Undergraduate GPA, MCAT Average, Number of Unacceptable Medical School Grades, NBME Parts 1 & 2, Comprehensive Clinical AssessmentPerformance, CBX Overall Scores and CBX Flag Totals (N = 159)

	GPA	*MCAT Average*	*Unacceptable Grades*	*NBME Part 1*	*CCA Score*	*NBME Part 2*	*CBX Ob/Gyn*	*Ob/Gyn Flags*	*CBX Surgery*
MCAT average	0.65**								
Unacceptable grades	–0.50**	–0.49**							
NBME Part 1	0.53**	0.67**	–0.56**						
CCA score	0.35**	0.37**	–0.43**	0.58**					
NBME Part 2	0.42**	0.54**	–0.44**	0.79**	0.70**				
CBX ob/gyn	0.02	–0.04	–0.05	0.001	0.09	0.12			
Ob/gyn flags	0.001	0.10	0.02	0.08	0.04	0.02	–0.36**		
CBX surgery	0.18*	0.29**	–0.26**	0.39**	0.42**	0.44**	0.33**	0.06	
Surgery flags	–0.17*	–0.31**	0.24**	–0.38**	–0.37**	–0.44**	–0.28**	0.08	–0.79**

*$\rho < .05$; **$\rho < .01$.

TABLE 2
Results of Multiple Regression Analyses to Predict Third-Year Medical Students' (N = 159) Performance on a Comprehensive Clinical Assessment and the NBME Part 2 Exam

Variable	b *Estimate*	*Standardized* β	t	p
Dependent variable = CCA score				
$R^2 = 0.40$, $\rho < 0.001$				
Intercept	41.64	0.00	5.39	<0.001
Undergraduate GPA	1.91	0.07	0.83	0.41
MCAT average	–0.71	–0.11	–1.13	0.26
Unacceptable grades	–1.02	–0.14	–1.74	0.08
NBME Part 1	0.04	0.45	4.79	< 0.001
Ob/gyn CBX score	0.04	0.005	0.07	0.95
Surgery CBX score	2.25	0.22	3.04	0.003
Dependent variable = NBME Part 2				
$R^2 = 0.65$, $\rho < 0.001$				
Intercept	105.66	0.00	8.72	<0.001
Undergraduate GPA	–0.15	–0.003	–0.04	0.97
MCAT average	0.49	0.04	0.50	0.62
Unacceptable grades	0.48	0.03	0.53	0.60
NBME Part 1	0.14	0.74	10.36	<0.001
Ob/gyn CBX score	1.64	0.09	1.70	0.09
Surgery CBX score	2.37	0.11	2.04	0.04

using NBME Part 2 pass/fail results as the gold standard measure of competency, that is, the measure by which NBME Part 1 and the CBX flag scores were assessed. As would be expected, given the relationship between the NBME exams, the NBME Part 1 exam is both sensitive (0.90) and specific (0.80). Between the two CBX measures, the ob/gyn flag score proved to have the greater sensitivity (0.86 vs. 0.60); that is, of the 154 students that passed the NBME Part 2 exam, 133 students had zero flags on the ob/gyn case, whereas only 92 students had zero flags on the surgery case. However, the surgery flag score had the greater specificity (1.00 vs. 0.20), that is, of the 5 students that failed the NBME Part 2 exam, all 5 students received one or more flags on the surgery case while only 1 student received one or more flags on the ob/gyn case.

In Table 4, sensitivity and specificity were calculated using CCA pass/fail results as the gold standard measure of competency. The NBME Part 1 again proved to be sensitive (0.90) but was less specific

TABLE 3

2 × 2 Tables for NBME Part 2 Performance by NBME Part 1 Performance, CBX Ob/Gyn Flag Score, and CBX Surgery Flag Score

	NBME Part 2					*False-Positive Rate*	*False-Negative Rate*
	Pass	*Fail*	*Total*	*Sensitivity*	*Specificity*		
NBME Part 1							
Pass	139	1	140	0.90	0.80	0.10	0.20
Fail	15	4	19	(139/154)	(4/5)	(15/154)	(1/5)
Total	154	5	159				
Ob/gyn flags							
None	133	4	137	0.86	0.20	0.14	0.80
1 or more	21	1	22	(133/154)	(1/5)	(21/154)	(4/5)
Total	154	5	159				
Surgery flags							
None	92	0	92	0.60	1.00	0.49	0.00
1 or more	62	5	67	(92/154)	(5/5)	(62/154)	(0/5)
Total	154	5	159				

TABLE 4

2 × 2 Tables for Clinical Assessment Performance by NBME Part 1 Performance, CBX Ob/Gyn Flag Score, and CBX Surgery Flag Score

	Comprehensive Clinical Assessment					*False-Positive Rate*	*False-Negative Rate*
	Pass	*Fail*	*Total*	*Sensitivity*	*Specificity*		
NBME Part 1							
Pass	141	3	144	0.90	0.57	0.10	0.43
Fail	15	4	19	(141/156)	(4/7)	(15/156)	(3/7)
Total	156	7	163				
Ob/gyn flags							
None	136	5	141	0.87	0.29	0.13	0.71
1 or more	20	2	22	(136/156)	(2/7)	(20/156)	(5/7)
Total	156	7	163				
Surgery flags							
None	93	2	95	0.60	0.71	0.40	0.29
1 or more	63	5	68	(93/156)	(5/7)	(63/156)	(2/7)
Total	156	7	163				

(0.57). As before, sensitivity favored the ob/gyn case, 0.87 versus 0.60 for the surgery case; whereas specificity favored the surgery case, 0.71 versus 0.29 for the ob/gyn case.

DISCUSSION

The results of the regression analyses indicate that the overall surgery CBX score and the NBME Part 1 score were the only significant predictors of CCA performance. The best predictor of CCA performance was the NBME Part 1 score. The overall surgery score was a better predictor of CCA performance than the overall ob/gyn score, undergraduate GPA, MCAT score, and number of unacceptable grades received during medical school. The overall ob/gyn score was not a significant contributor to the prediction of the CCA score and was not a better predictor than the other measures.

In the prediction of NBME Part 2 performance, the only significant predictors were the overall surgery score and the NBME Part 1 score. The NBME Part 1 score was by far the best predictor with a β coefficient of .74 compared to a β coefficient of .11 for the overall surgery score. Although the importance of the ob/gyn score was higher in this model, it still lacked statistical significance, $p = 0.09$. (In a regression model that excluded the surgery score, the ob/gyn score became a significant predictor of NBME Part 2. In the model using both scores, any shared contribution in the prediction of the NBME Part 2 score was attributed to the stronger surgery score.)

The results of the sensitivity and specificity analyses also indicate a difference in the two CBX cases. The ob/gyn flag score had greater sensitivity (more accuracy in passing competent students), whereas the surgery flag score had greater specificity (more accuracy in failing noncompetent students).

On average, the students performed differently on the two computerized clinical problems. In the present study, it is likely that more students would have been exposed to patients similar to the ob/gyn case than to the surgery case. More experience with ob/gyn patients would reasonably explain some of the difference in performance. Nonetheless, the surgery case appeared to be more difficult than the ob/gyn case, as can be seen in the number of flags (or critical management mistakes) made on each case. If the clinical competence assessment was designed to measure accurately what third-year clerks are expected to have learned by the end of the academic year, then one of two implications follow: Either many students have not mastered what they were expected to have mastered (i.e., the student's knowledge

level is inadequate), or not all students had the learning experiences the faculty thought they should have had (i.e., the curriculum is remiss). The source of the problem is difficult to identify and is likely to be a combination of both factors rather than either one exclusively. These findings, however, show that feedback on performance is critical, both in designing remediation for individual students and in revising the curriculum (or faculty expectations for student performance) so that a greater degree of mastery is achieved.

What is interesting, and can be gleaned only by examining both the results from the Bayesian analysis of test characteristics for each case and the classical analysis, is the fact that the higher difficulty of the surgery case contributed to increased variance in student performance. Therefore, the surgery case had higher validity coefficients with other performance measures and higher predictability for the measures of overall clinical performance. However, this was accomplished at the expense of increasing its false positive rate (failing students who should pass) and thereby decreasing the case's sensitivity, perhaps to unacceptable levels. The opposite was true for the ob/gyn case. The sensitivity and false-positive rates were better, although its specificity and true-positive rates were worse than those of the surgery case. In essence, if one is to use these two CBX cases, one is confronted with making explicit the trade-offs, the acceptable types of classification errors, and the magnitudes of the classification errors for student performance that an institution is willing to accept (Algina, 1978; Millman, 1989; Wolf, 1991). Another important point is that student performance on the ob/gyn case indicated a high degree of competence, something the curriculum was designed to accomplish and the clinical competence exam was designed to assess. This high degree of competence, in turn, attenuated the variance in performance and thus the classical measures of validity for this case. As such, curriculum and assessment goals and needs (which are more mastery and criterion-referenced) conflict with the norm-referenced foundation underlying classical test and measurement theory. This suggests that test characteristics derived from a detailed Bayesian analysis of 2×2 tables may be equally, if not more, appropriate for assessing and understanding clinical-based assessment measures than traditional psychometric analyses.

Clearly, more detailed study is needed as a follow-up to this investigation. The false-negative and specificity rates in the present study were based on very small sample sizes as few students failed in their performance. Unfortunately, CBXs were not included in subsequent CCAs at the UMMS. Aggregating results across successive classes (years) within one school and/or among multiple schools would be a helpful next step in characterizing the implications of differential case characteristics. It is also important to be creative in developing measures of clinical performance with better fidelity and consistent with actual medical practice that can be used as gold standards in future studies.

IMPLICATIONS

Given that the cost of testing students with computer simulations of patient problems is higher than paper and pencil tests, it would seem logical to use the computer examinations only if they provided information about student performance that could not be measured with a conventional MCQ test. The evidence in this study is both encouraging and discouraging. The two cases sampled performed very differently in terms of sensitivity and specificity. The surgery case, although perfectly specific (all of the students who failed the NBME Part 2 received one or more flags on the surgery case), was not very sensitive (62 of 154 students who passed Part 2 received one or more flags). The ob/gyn case was more sensitive than the surgery case (only 21 of the 154 students who passed Part 2 received one or more flags) but not specific (4 of 5 students who failed Part 2 received no flags on the ob/gyn case). The CBX cases may, however, more closely approximate the real world of medicine and may, because of the increased fidelity, be better measures of what students will actually do in practice than traditional MCQ tests. Additional validation studies may include comparing CBX performance to performance on standardized patients and to observed patient presentations and interactions.

One of the goals of the CCA at the UMMS was to adopt a realistic examination setting that would place the student in an environment that more closely resembles the practice environment. Although clinical examinations, and for that matter computer simulations, approxi-

mate actual practice behaviors, the questions of reliability, validity, sensitivity, and specificity of the individual components of the exam need to be determined before these types of examinations are suggested for wider use.

In his analysis of the challenges inherent in evaluating the competence for professional practice, McGaghie concludes that "Results from the competence assessments need to be interpreted and used with an understanding of the limits (not failure) of current assessment technologies" (McGaghie, 1993, p. 243). Based on the data presented in this study, the limitations of the CBX exam and of using only two cases to assess student competence are clear. Our data do not support wide-scale substitution of CBX cases for actual patient interactions, and the use of only two cases cannot be defended as an adequate sample of student performance. However, the results do indicate that CBX-type cases can complement and augment other forms of clinical educational experiences. The results also shed light on the impact individual cases can have on error rates and the necessity of making explicit the trade-offs required in true and false positive rates when setting passing scores and competency criteria (e.g., Algina, 1978).

FUTURE STUDIES

The application of sensitivity and specificity analysis to tests of student competence has appeal and was demonstrated in this pilot study to assist in determining the usefulness of the two CBX cases. The use of these traditionally epidemiological indexes may be appropriate for assessing the usefulness of patient simulations, clinical examinations, and test stations in an objective structured clinical examination to determine if the examination is achieving the desired outcome. Hopefully, gold standards other than the NBME Part 2 examination will be identified and used in such studies. Such gold standards may include comparisons of student performance on similar patient problems presented in a simulation, by a standardized patient, and in an actual doctor-patient encounter. Such comparisons would assist in determining the limits and potential of these simulation techniques. Although it is possible to develop parallel CBX and

standardized patient encounters (Norman & Feightner, 1981), it is more difficult to equate real patient presentations to simulations.

Basically, this study is a small part of a long tradition of validity research aimed at selecting assessment problems and strategies that will assist in identifying competent practitioners and in identifying weaknesses in student preparation. With further refinement, the assessment of clinical performance may be achieved through testing formats other than the conventional multiple choice question exam. Particularly helpful would be a test format that more closely approximates the complex patient-physician encounter. The CBX-type formats may be a step in that direction.

REFERENCES

Algina, J. (1978). On the validity of examinations for making promotion decisions in medical education. *Medical Education, 12*, 82-87.

Clyman, S. G., & Orr, N. A. (1990). Status report on the NBME's computer-based testing. *Academic Medicine, 65*, 235-241.

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

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving: A ten-year retrospective. *Evaluation and the Health Professions, 13*(1), 5-36.

Freund, R. J., Littell, R. C., & Spector, P. C. (1986). *SAS system for linear models*. Cary, NC: SAS Institute.

McGaghie, W. C. (1993). Evaluating competence for professional practice. In L. Curry & J. F. Wergin (Eds.), *Educating professionals* (pp. 229-261). San Francisco, CA: Jossey-Bass.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5-9.

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

Swanson, D. B., & Stillman, P. L. (1990). Use of standardized patients for teaching and assessing clinical skills. *Evaluation and the Health Professions, 13*(1), 79-103.

van der Vleuten, C.P.M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*(2), 58-76.

Volle, R. L. (1990). Standardized testing of patient management skill: A computer-based method. *Clinical Orthopaedics and Related Research, 257*, 47-51.

Wolf, F. M. (1991). Toward a theory of clinical evidence: An empirical scientific perspective. *Professions Education Researcher Quarterly, 12*(4), 3-7.