An Action Research Approach to Standardizing the Evaluation of Diagnostic Psychomotor Skills

David P. Waalen, B.A., D.C., Canadian Memorial Chiropractic College, **Judith K. Waalen, Ph.D.,** Ryerson Polytechnic University, and **Franklin J. Medio, Ph.D.,** Medical University of South Carolina

The effective evaluation of performance diagnostic skills is essential to determine the clinical competency of students in the health care professions. This study describes an iterative, participatory approach to the development of standardized methods for the evaluation of students' diagnostic psychomotor skills. An action research design was utilized to foster a collaborative appraisal of current performance assessment practices in the context of a faculty development program. This program provided a catalyst for a consensual process that established appropriate performance criteria and developed new evaluation instruments. Statistical comparisons were made between the new and the old assessment instruments with respect to examiner variability. All faculty ($n = 10$) involved in evaluating diagnostic psychomotor skills, and all students ($n = 147$) enrolled in the Introductory Diagnosis course were utilized in these comparisons. When compared to the original evaluation instruments (F ratio = 13.69, 9 df, $p = 0001$), variability among evaluators by instructor group was reduced (F ratio = 2.43, 9 df, $p = 01$) with the new instruments. Post-hoc significant differences between group means (using the Tukey B) dropped from 20 to only 1 difference. More consistent evaluation of diagnostic psychomotor skills can be accomplished by: clearly defining performance criteria, designing appropriate evaluation instruments, and establishing an iterative process of instruction and feedback for faculty evaluators. (The Journal of Chiropractic Education 14(2): 78–87, 2000)

Key words: clinical competence, diagnostic skills, evaluation, examiner variability

INTRODUCTION

Educators of health care professionals have the challenging responsibility of ensuring that their students acquire the comprehensive diagnostic skills and knowledge required to competently conduct their subsequent practice. This diagnostic knowledge base has been evaluated in many ways: paper-and-pencil tests, performance exams, and direct observation in a clinical setting. There has been increasing dissatisfaction with the use of written tests, since the diagnostic skills of health care professionals are not applied this way in actual clinical practice (1). Although performance evaluations are often more costly and difficult to administer than written tests, there is growing recognition among authorities on competency assessment of the need to employ these more "real-to-life" testing procedures (2). However, the accurate assessment of clinical psychomotor skills can be compromised by varying faculty standards of students' performance (3), by the design of the evaluation instruments used (4), and by the type of training provided to the evaluators (5).

The purpose of this study was fourfold: to educate examiners on the importance of conducting standardized performance evaluations; to design, by consensus, new evaluation forms to assess students' performance of various diagnostic skills; to conduct performance evaluations using the new assessment

The Journal of Chiropractic Education

Vol. 14, No. 2. Printed in U.S.A.
1042-5055/$4.00

instruments; and to compare the results from the new evaluation instruments to those from the ones previously used in order to assess any improvement in examiner variability.

Written recognition and recall tests have generally been found to be objective, reliable, and relatively easy to administer, but they are often poor indicators of a student's clinical abilities (6). The acquisition of factual knowledge alone, although an important component of clinical competency, cannot be equated with the diagnostic performance and problem-solving skills that are the essential attributes of a capable practitioner (7). While some paper-and-pencil tests may have the ability to indicate outcomes on other written tests, the relationship between results from written examinations which purportedly predict clinical competency and those from objectively scored performance evaluations has been shown to be weak (8). Some investigators (9) have found that results from written tests on knowledge of skills correlate well with results from performance evaluations of more senior students; however, the correlation between written and performance test results was weak for students at the "beginner stage" of the competency continuum (10,11). Many professional licensing bodies, such as the Medical Council of Canada, have confirmed the importance of adequately assessing the performance of clinical skills in their candidates for licensure by incorporating performance evaluation components into their testing procedures (12).

LITERATURE REVIEW

Especially in the evaluation of diagnostic skills, performance exams are critical to establishing clinical competency and cannot be replaced by recall and recognition methods of testing (7). However, the development of a standardized method of assessing psychomotor performance faces three challenges. First, it is difficult to specify the behavioral attributes for evaluation. Typically, a skill is either adequately executed or it is not. The second difficulty involves standardizing the performance assessments. Many faculty evaluators feel that they are well qualified, simply by virtue of their years of clinical experience, to make these assessments with respect to what constitutes the competent and complete performance of a particular diagnostic psychomotor skill. The third challenge involves the multiplicity of examiners who bring their own intuitive assessment approaches to the skill set being examined. Since the capacity to perform these psychomotor skills is the quintessence of clinical competence, a meaningful evaluation of them demands a focused, consistent method of assessment that limits variability among different examiners. Much of the current interest in performance assessment stems from concerns that traditional testing does not lead to curriculum renewal or improved student performance and not from concerns about better measurement for its own sake (13). Some hold the view that quality control is an empirical matter and needs to involve the measurement of reliability, validity, and objectivity (8,14,15).

In our review of the literature, three themes emerged: the discomfort with relying on written examinations to measure psychomotor performance, examiner variability in rating performance, and the challenge of defining clinical competence.

Written Examinations Versus Performance Assessments

There has been growing dissatisfaction with the use of multiple-choice testing because the knowledge of health care professionals is not challenged this way in a clinical environment (1). Although performance-based testing is more costly, the active (rather than passive) behavior that is required from the students provides a more valid predictor of their performance in clinical practice.

In a study conducted by Barrows et al. (6), trained simulated patients were used to assess the clinical competence of senior medical students in a multiple-station performance examination. They attempted to deal with the concern of residency directors that although some medical school graduates score highly on written examinations, they may nevertheless be clinically inept. They developed and tested a performance assessment that identified strengths and deficiencies in students' diagnostic skills. After the assessment, feedback was provided to the students that delineated performance expectations for subsequent assessments. Since time and expense can limit the feasibility of this approach (16), most health training institutions use a combination of written and performance-based testing. Ginsburg (17) studied the differences in skill rating produced by faculty evaluators and students' performance on paper-and-pencil tests. To his dismay, the correlation between them was slight and the correlation between two

skill assessments was negligible. Although inter-rater variability was not measured in this study, the author suggested that the faculty needed better training in evaluative skills.

Since the practice of chiropractic has a strong psychomotor component (18), the reliance on paper-and-pencil measures of performance is particularly inappropriate for chiropractic students. Most beginning students have no manual therapy or clinical examination skills but must develop them in order to ultimately function as neuromusculoskeletal specialists. In a study by Waalen and Waalen (5), the system used by chiropractic clinicians to evaluate clinical competency compared scores on objective structured clinical examinations (OSCE) and academic grades. Only weak to moderate relationships were found. Suggestions were made to empirically measure the evaluation instrument used and to spend time training clinicians in its appropriate use.

Examiner Variability

A number of researchers have concerned themselves with specifying evaluation standards in order to improve low inter-rater reliability (2,19–21). In one study, for example, the level of inter-rater agreement was examined on various competencies and behaviors considered to be important in pediatric residents (22). Rating applicants for professional certification is required by the licensing body, so verification of clinical competence from their program directors is essential. Their findings suggested that faculty assessors had poor inter-rater agreement (Kappa values less than .40) on most of the ratings, but particularly on assessing items linked to relating to patients, staff, and colleagues. There was no suggestion that assessors were involved in the development of the assessment tool or specifically trained to use it.

Carline et al. (23) examined factors affecting the reliability of rating clinical skills, and suggested that either evaluators are unreliable or they are observing different types of performance in the same student. Using a nine-item instrument, evaluators rated students using a four-point scale (1 = unsatisfactory, 4 = excellent). Several ratings were done for each scale. Although the reliability on some of the items was good, some aspects of clinical performance (relationships with patients) were poor. There was no mention of a training component for the faculty members doing the assessment; in fact, the authors suggested that more careful training might improve the reliability of ratings.

Some researchers have used instruments developed to measure goals and objectives of the educational program. Others have developed models of assessment in conjunction with those who will actually be using it. In a study by Fiel et al. (24), a model for evaluating psychomotor skills was developed for osteopathic students. Consensus on the model was achieved by engaging the evaluators in discussions until unanimous agreement was reached. Specifically, the evaluators discussed each step in the skill being examined and each criterion used in the rating system. The authors attribute their success to the intimate involvement of each evaluator in the development of the model.

Measuring Clinical Competence

The literature in this area suggests that competency-based education is best achieved when evaluation procedures have fidelity to real-world experience (7), but little attention has been devoted to the direct assessment of professional skills in a realistic environment (25). Well-designed rating forms provide one way of evaluating clinical performance when they contain a clear perception of the goals of the course or program (4). The way in which performance indicators have been developed varies. Lane et al. (7) reported on a strategy involving a 2-day meeting that included plenary, large-group, and small-group sessions. Their task was to define what preventive medicine residents should be able to do. This developmental process involved the production of draft documents triggered by discussions and a final handbook of evaluation strategies.

In Bramble's study (26), clinical competence was measured using an eight-part clinical skill tool. The rating scale ranged from 0 (poor performance) to 4 (strong performance). This tool was derived from the goals and objectives of the nursing program and was used to compare subjects who had objective structured clinical assessment (OSCA) training and those who did not. Raters were requested to use it but had no input into its development.

Chambers and Glassman (11) focused on a five-stage developmental model of clinical expertise (novice, beginner, competent, proficient, and expert) when discussing performance assessment. In their view, an attempt should be made to consider where the student is on the professional growth curve when designing assessment instruments. This model provided the

guidelines for evaluating skill performance used in this study since it has been demonstrated that in beginning courses, different competencies are relatively independent from one another. As education and proficiency increases, competencies tend to evolve and integrate (9).

Our review of the literature indicated that few studies empirically examined the instruments being used to measure the diagnostic psychomotor performance of students. Further, some authors reported that training of evaluators in the use of the evaluation instruments was not done despite recommendations to the contrary. Finally, although in some studies critical reflection on the need for examiner consistency was present, few systematically examined ways to achieve it. Therefore, the authors of this study chose an action research model to examine the assessment of student diagnostic psychomotor skills. Action research is a methodology that pursues change and research at the same time by using a cyclical process alternating between action and critical reflection. It is a highly participatory and iterative process since it is felt that beneficial change is best achieved when those affected by it are closely involved (27).

THE CONTEXT

The Introductory Diagnosis course at the Canadian Memorial Chiropractic College is taught during the 2nd year of the 4th-year program. It is presented after the students have taken the required basic science courses such as anatomy and physiology, and before they advance to more specialized diagnosis courses such as orthopedics, neurodiagnosis, and psychology in their 3rd and 4th years of study. The course is designed to provide students with the basic diagnostic information and physical examination skills that are essential to a primary contact health care provider. This is accomplished by lecture format instruction, followed by the demonstration and supervised practice of the relevant physical examination procedures in an appropriately equipped, small-group, clinical diagnosis lab setting. Each group of approximately eight students is supervised by an experienced faculty instructor who provides guidance, feedback, and coaching to the individual students in the group as they perform the various components of a particular type of physical examination on one another. For instance, the instructor ensures that all aspects of an abdominal examination—inspection, auscultation, percussion, palpation, and special maneuvers—are capably performed by each of the students in the group.

Although three multistation OSCE evaluations are utilized to test the competency of chiropractic students in later years of the program, time and budgetary constraints preclude their use in the 2nd-year Introductory Diagnosis course. A written multiple-choice examination and two performance evaluations therefore primarily achieve the evaluation of the skills and knowledge that a student learns in the course. The instructors near the end of the fall and spring terms conduct these performance evaluations. A student is randomly assigned particular physical examinations (such as an eye exam, a cardiovascular exam, or an abdominal exam) to carry out on a fellow student. Instructors grade the student's performance using a checklist prepared for the specific examinations that have been assigned. The performance evaluations account for 30% of the final mark and the student must pass each performance evaluation to successfully complete the course. Since these performance evaluations are essentially subjective judgements on the part of faculty evaluators, each of whom has a somewhat different perspective on what constitutes the acceptable performance of a particular examination procedure, interexaminer variability has been a long-standing issue with this evaluation component of the course.

In an attempt to standardize the methods used to evaluate the diagnostic performance skills of students in the Introductory Diagnosis course, evaluation procedures had been modified in an "ad hoc" manner over several years with no systematic input from faculty evaluators, and had gradually evolved from rather simplistic and highly subjective global appraisals (acceptable or unacceptable) to complex common evaluation forms with multilevel rating scales for different aspects of each type of physical exam being evaluated. The evaluation methodology that emerged from this process was thought to adequately control the subjective variability in the marking of the 10 different instructors in the course. However, informal feedback from some students and faculty members suggested that this supposition might be erroneous, and, since no thorough appraisal of the performance evaluation methods had ever been carried out to substantiate it, a faculty development program for the instructor evaluators was convened to foster critical reflection

on issues related to the evaluation of performance skills.

METHODS

The faculty development program on performance evaluation methods was facilitated by an educational psychologist with extensive experience in health care education. The program emphasized that the evaluation of psychomotor skills was an essential element in the assessment of a student's overall diagnostic ability that could not be replaced by written examinations or other testing methods which had low fidelity to practice reality. It also stressed the importance of establishing standards and evaluation instruments that were consistent with the students' current position on the Chambers five-stage competency continuum (11). It was pointed out to the faculty evaluators that the clearly discernible differences in performance expectations between a 2nd-year "beginner" and a "competent" 4th-year student about to graduate must be given careful consideration when selecting a suitable evaluation strategy.

With these precepts in mind, the educational psychologist conducted a facilitated discussion among the faculty evaluators in order to identify weaknesses in the existing performance assessment methods. This was accomplished, in part, by the use of a videotape which depicted an "actor" student conducting an abdominal examination. Faculty participants were then asked to evaluate the competency of the student using the existing instrument. The critical appraisal that followed concluded that the existing performance evaluation instruments (see Fig. 1 for an example), which had been adapted from those used to evaluate the performance of 4th-year students in the chiropractic outpatient clinics, were inappropriate for students at the beginner stage of diagnostic skill development. It was also felt that the current evaluation grading forms were too complex and difficult to score, and therefore were not conducive to controlling subjective variability among the 10 different faculty evaluators. It was therefore decided to thoroughly analyze the results from the existing evaluation instruments to determine the extent to which variability in grading among the faculty evaluators was, in fact, problematic.

Each of the 147 students in the Introductory Diagnosis course was randomly assigned to one of the 10 faculty instructors. Every instructor was responsible for two separate groups of seven or eight students for a total of approximately 15 students per instructor. At the end of the fall term, individual students had their diagnostic performance skills evaluated by their instructor using the standardized checklists that had been developed for the various physical examination procedures. A one-way analysis of variance was conducted comparing the mean scores from the 10 different instructor/student groups. The results indicated that the current performance evaluation methodology did not adequately control subjective variability in grading among the 10 instructors and that further work needed to be done in this regard. This analysis and the facilitated discussions in the faculty development program laid the foundation for a consensus process which designated the principal elements of the various examination procedures and their corresponding rating scales which were subsequently incorporated in the new evaluation instruments (see Fig. 2 for an example). After specific instruction as to their use, the new performance evaluation checklists that emerged were employed by the same faculty members to evaluate the same groups of students at the end of the spring term that they had evaluated in the previous term. Feedback on their Time 1 ratings was also provided to the evaluators by using a graphic representation of the post hoc comparisons (Tukey B).

RESULTS

To determine whether the new performance evaluation instruments, produced as a result of the consensual faculty development program, reduced variability among the 10 different faculty examiners, a one-way analysis of variance between the mean scores from the 10 different instructor groups was carried out and the results compared to those from the original instrument. The analysis of variance conducted on the data using the original instrument yielded an F ratio of 13.69, 9 df, $p = .0001$ and the Tukey B post hoc test revealed that there were 20 significant differences between the mean scores. Four instructor groups had mean differences that were either above or below the 95% confidence interval for the overall mean (see Table 1). The analysis of variance carried out on the data using the new instrument yielded an F ratio of 2.43, 9 df, $p = .01$ and the Tukey B post hoc test revealed that only one instructor group had a mean difference that was significantly different from that of one other instructor group. No group was

INSPECTION:	
Approach:	Promote patient relaxation
	Tangential light, head and knees up, bladder empty
	Scars, dilated veins, hernias, pulsations, masses
AUSCULTATION:	
Bowel sounds:	5–34 per min (absent > 2 min)
Bruits:	Aortic, renal, iliac, femoral
PERCUSSION:	
Stomach:	Tympanic percussion note
Liver:	Dullness (6–12 cm MC, 4–8 cm MS)
Splenic perc.	
Sign:	Last IC space L. ant. axillary
PALPATION:	
Light:	Fingerpads—tenderness, resistance
Deep:	Possible 2-hand technique—4 quadrants
	Liver—sandwich and hooking techniques
	Gall bladder—Murphy's sign
	Kidneys—R. capturing technique—punch test
	Aorta—lateral pulsations (> 2 cm)
SPECIAL MANEUVERS:	
Ascites:	Percussion, shifting dullness, fluid wave
Appendicitis:	McBurney's point, hyperesthesia, rebound tenderness, Rovsing's sign, psoas and obturator signs
DDxc:	Abdominal wall mass from intra-abdominal mass

Evaluation	Inadequate	Satisfactory	Very Good	Exceptional	Marks
Relates to patient in effective, empathetic & professional manner	_____	_____	_____	_____	_____
Organizes examination systematically and efficienctly	_____	_____	_____	_____	_____
Performs a thorough and complete examination	_____	_____	_____	_____	_____
Understand area being examined reasons for exam procedures	_____	_____	_____	_____	_____
Demonstrates skillful use of diagnostic instruments/techniques	_____	_____	_____	_____	_____

Figure 1. An example of the original rating system for an abdominal examination.

either completely above or below the 95% confidence interval for the overall mean (Table 2). A paired t test demonstrated that there was no significant difference between overall mean scores using the original and the new instrument ($t = -.06$, $p > .05$).

An analysis of variance by instructor group was also conducted on the data from the students' final written exam. No significant differences were found between the mean scores of the 10 instructor groups (F ratio = .37, 9 df, $p = .95$). The relationship between the students' marks on the second performance evaluation and their marks on the final written examination was found to be weak ($r = 29$, $p = .0001$).

Abdominal Examination	Inadequate	Acceptable	Well Done
Inspection			
Patient position	___	___	___
Abnormalities	___	___	___
Auscultation			
Bowel sounds	___	___	___
Bruits	___	___	___
Percussion			
Stomach	___	___	___
Liver	___	___	___
Spleen	___	___	___
Palpation			
Liver & gallbladder	___	___	___
Kidneys	___	___	___
Aorta	___	___	___
Special maneuvers			
Appendicitis	___	___	___
Ascites	___	___	___
Empathy & professionalism	___	___	___
Organization & efficiency	___	___	___
Knowledge & understanding	___	___	___

Figure 2. An example of the revised rating system for an abdominal exam.

Table 1. Means, Standard Deviations, and Confidence Intervals for Original Instrument

Group	*N*	Mean	SD	SE	95% CI
1	15	12.96	1.34	0.35	12.22–13.70
2	16	10.24	1.40	0.35	09.50–10.99
3	15	13.73	1.58	0.41	12.85–14.60
4	14	13.32	0.64	0.17	12.95–13.69
5	14	11.64	1.53	0.41	10.76–12.52
6	16	12.73	0.67	0.17	12.37–13.08
7	16	13.73	1.11	0.28	13.14–14.31
8	16	12.96	1.02	0.26	12.42–13.51
9	14	11.19	1.40	0.37	10.38–12.00
10	11	12.22	0.95	0.29	11.58–12.86
Total	147	12.49	1.61	0.13	12.22–12.75

F Ratio = 13.69, 9 df, $p = .00001$.

DISCUSSION

For cogent and compelling reasons, the use of competency-based evaluation methods has become ubiquitous in professional health education programs, and a wide variety of evaluation strategies have been developed in an attempt to assess the diagnostic skills and knowledge of students. This diversity has provoked an ongoing discussion in the literature with respect to the accuracy, utility, and appropriateness of many of the evaluation methods that have been employed. However, there appears to be a growing consensus among health care educators that performance is an essential element of competency which is best evaluated by methods that have high fidelity to actual clinical situations (1,2). Support for the contention that performance is a distinct dimension of diagnostic competency can be found in previous

Table 2. Means, Standard Deviations, and Confidence Intervals for Revised Instrument

Group	*N*	Mean	SD	SE	95% CI
1	16	12.44	1.62	0.41	11.58–13.31
2	16	12.46	1.34	0.34	11.74–13.17
3	14	13.86	1.29	0.34	13.11–14.60
4	14	12.84	1.43	0.38	12.01–13.66
5	14	12.56	1.47	0.39	11.71–13.40
6	16	13.18	1.08	0.28	12.61–13.76
7	16	13.03	1.05	0.26	12.47–13.58
8	16	11.94	1.42	0.36	11.19–12.70
9	14	12.79	1.01	0.27	12.20–13.37
10	11	12.22	1.39	0.42	11.29–13.15
Total	147	12.73	1.38	0.11	12.51–12.96

F Ratio = 2.43, 9 df, $p = .01$.

research which disclosed disappointing relationships between results on written examinations and those from performance evaluations (5,9,17). Our study, which found a weak correlation between students' marks on the written examination and their scores on the performance evaluation, lends further credence to this contention.

Evaluating the performance component of diagnostic competency presents a complex and multifaceted challenge to health care educators. Common performance criteria must be agreed upon by the evaluators (19), and the expected level of student performance must be specified (2). Evaluation instruments must be designed to minimize variability in scoring among evaluators and they must take into account the students' position on the competency continuum (9,11). An iterative process of evaluator training and feedback with respect to the use of the instruments is also essential to reduce variability among the faculty raters.

This study took the issues that had been raised in the literature into consideration when redesigning the performance evaluation strategy for the Introductory Diagnosis course. The faculty development program for evaluators emphasized the importance of the performance aspect of competency and facilitated the consensual process that was used to establish performance criteria that were appropriate for students at the "beginner" stage of clinical competency. The critical appraisal of the original evaluation instruments conducted in the faculty development program identified several inadequacies, and the results of the analysis of variance demonstrated their ineffectiveness in controlling variability as well as providing a benchmark for assessing the new evaluation instruments that were the outcome of the action research process. Although variability among evaluators was not completely eliminated with the new instruments (probably an unrealistic goal with an essentially subjective evaluation), our findings indicate that, by using an action research approach, it is certainly possible to reduce examiner variability.

Although the action research approach to decreasing examiner variability that we have described in this study has demonstrated a measure of success, some general caveats are worthy of commentary. It must be emphasized that performance is only one component, albeit an important one, of clinical competency, and any comprehensive assessment strategy must certainly address issues such as cognition and integration as well as performance. The validity and reliability of any testing procedure is another obvious concern for faculty evaluators. However, these topics are part of a much larger discussion that is beyond the scope of this paper.

The inherent limitations imposed by the multidimensional nature of action research must also be acknowledged. While several factors such as faculty education, feedback, and participation in the process, as well as the use of new evaluation instruments probably contributed to the favorable results, it is not possible to determine the extent to which any single factor was responsible for the reduction in examiner variability.

Perhaps the most daunting deterrent to conducting this type of action research relates to the demands it makes upon the all too frequently rather limited resources of chiropractic educators. This trenchant, time-consuming, and often contentious process requires a very high level of commitment from faculty

members and academic administrators alike. Action research is a dynamic, ongoing, iterative process that requires periodic faculty re-evaluation, re-education, and feedback. As modifications are introduced to the curriculum, and as new, less-experienced faculty evaluators join the team, it is essential to reinstitute the process to ensure a minimum of variability among faculty evaluators. Despite its many demands, the authors of this study feel that the action research approach can play an important role in standardizing evaluations of the performance elements of clinical competency.

CONCLUSION

This study stressed that the evaluation of diagnostic psychomotor skills is an essential element in the overall assessment of students' clinical competency. It described a multifaceted action research approach to standardizing the evaluation of these important performance skills. This study demonstrated that this approach, which included clearly defining performance criteria, designing appropriate evaluation instruments, and establishing an iterative process of faculty instruction and feedback, can effectively limit variability in scoring among faculty members who assess these critical clinical skills.

ACKNOWLEDGMENTS

The authors of this study wish to acknowledge the contributions of Brian Schut, D.C., Associate Clinical Professor and Chair of the Clinical Diagnosis Department at CMCC, and the faculty of the Introductory Diagnosis course, without whose participation and commitment to improving performance evaluation, this project would not have been possible.

Received, April 13, 1999
Revised, March 15, 2000
Accepted, April 15, 2000
Reprint requests: David P. Waalen, Canadian Memorial Chiropractic College, 1900 Bayview Ave., Toronto, Ontario, M4G 3E6, Canada

REFERENCES

1. Haydon III R, Donnelly M, Schwartz R, Strodel W, Jones R. Use of standardized patients to identify deficits in students performance and curriculum effectiveness. Am J Surg 1994;168:57–65.
2. Newble D, Dawson B, Dauphinee D, Page G, Macdonald M, Swanson D, Mulholland H, Thomson A, Van der Vleuten C. Guidelines for assessing clinical competence. Teach Learn Med 1994;6(3):213–220.
3. Ainsworth MA, Speer AJ, Solomon DJ. A clinical evaluation form to improve faculty critique of students. Acad Med 1995;70:445.
4. Gugelchuk GM, Daum SG. Guidelines for designing resident rating forms. J Am Osteop Assoc 1992;92(6): 787–794.
5. Waalen DP, Waalen JK. A comparison of clinical competency evaluation methods. Chiro Ed 1996;9(4): 147–154.
6. Barrows HS, Williams RG, Moy RH. A comprehensive performance-based assessment of fourth-year students' clinical skills. J Med Educ 1987;62:805–809.
7. Lane DS, Parkinson MD, Ross V, Chen DW. Performance indicators for assessing competencies of preventive medicine residents. Am J Prevent Med 1995; 11(1):1–8.
8. Mitchell KJ. Traditional predictors of performance in medical school. Acad Med 1990;65:149–158.
9. Van der Vleuten CPM, Van Luyk SJ, Beckers HJM. A written test as an alternative to performance testing. Med Educ 1989;23:97–107.
10. Chambers DW. Some issues in problem-based learning. J Dent Educ 1995;59(5):567–572.
11. Chambers DW, Glassman P. A primer on competency-based evaluation. J Dent Educ 1997;61(8):651–666.
12. Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. Adv Health Sci Educ 1997;2:201–211.
13. Dunbar SB, Koretz DM, Hoover HD. Quality control in the development and use of performance assessments. Appl Measure Educ4 1991;(4):289–303.
14. Norman GS, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Med Educ 1991;25:119–126.
15. Paul VK. Assessment of clinical competence of undergraduate medical students. Indian J Pediatr 1994;61: 145–151.
16. Elliot DL, Fields SA, Keenen TL, Jaffe AC, Toffler WL. Use of a group objective structured clinical examination with first-year medical students. Acad Med 1994;69(12):990–992.
17. Ginsburg AD. Comparison of intraining evaluation with tests of clinical ability in medical students. J Med Educ 1985;60:29–36.
18. Good CJ. Defining quality in chiropractic education.J Chiropr Educ 1995;9(1):27–38.
19. Stiggins RJ. Design and development of performance assessments.Educational Measurement: Issues and Practice. Washington, DC: National Council on Measurement in Education. 1987;33–42.
20. Roberts J, Norman G. Reliabiity and learning from the objective structured clinical examination. Med Educ 1990;24:219–223.
21. Rothman AI. Understanding the objective structured clinical examination. Aust N Z J Surg 1995;65:302–303.
22. Davis JK, Inamdar S, Stone RK. Interrater agreement and predictive validity of faculty ratings of pediatric residents. J Med Educ 1986;61:901–905.

23. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med 1992;7:506–510.
24. Fiel NJ, Griffin RE, McNeil JA, Ajunwa MI, Salisbury CA, Aasved C. A model for evaluating student clinical psychomotor skills. J Med Educ 1979;54:511–513.
25. McGaghie WC. Professional competence evaluation. Educ Res 1991;20(1):3–9.
26. Bramble K. Nurse practitioner education: enhancing performance through the use of the objective structured clinical assessment. J Nurs Educ 1994;33(2):59–65.
27. Dick R. What is action research? http://www.scu.edu.au/schools/sawd/ari/whatisar.html, 1999.