



สารศิริราช

SIRIRAJ HOSPITAL GAZETTE

จัดพิมพ์โดยอนุमतคณะกรรมการคณะแพทยศาสตร์ศิริราชพยาบาล
Published Under the Auspices of the Faculty of Medicine, Siriraj Hospital

ปีที่ 56, ฉบับที่ 7, กรกฎาคม 2547

Volume 56, Number 7, July 2004

The Validity and Reliability of Clinical Performance Assessment of Residents in the General Surgery Residency Program

Cherdsak Iramaneerat, M.D., M.H.P.E.*

Abstract : A retrospective study of clinical performance assessment of residents in the general surgery residency program at the Department of Surgery, Faculty of Medicine Siriraj Hospital, during the 2000 - 2001 and 2001 - 2002 academic years was done to evaluate the inter-rater reliability, internal structure, and educational discriminant validity of the assessment. The inter-rater reliabilities were determined by using the intraclass correlation. Correlations between individual performance ratings were checked to determine how well faculty members differentiate their ratings with individual clinical skills. Finally, the percentage of marginal or unsatisfactory ratings was checked to demonstrate how sensitive the ratings were in identifying performance deficiencies among residents.

From the evaluation of 42 residents in the 2000 - 2001 academic year, inter-rater reliability coefficients of individual ratings ranged from 0.06 to 0.93 with an average of 0.51. From the evaluation of 47 residents in the 2001 - 2002 academic year, inter-rater reliability coefficients ranged from 0.04 to 0.85 with an average of 0.49. Inter-rater reliabilities of performance ratings of third-year residents were at an acceptable standard for a medium-stake assessment. However, those of first and second-year residents should need some improvement. Tasks performed by first and second-year residents seem to have been inadequate for faculty members to evaluate their performance reliably. The correlation study between individual performance ratings demonstrated the lack of discrimination between many clinical skills in faculty ratings. Faculty members need some guidelines to differentiate between many items including relationship with patients, relationship with other doctors, and relationship with other workers; knowledge and clinical judgment; work concentration and work effectiveness; and punctuality and responsibility. About ten percent of ratings were marginal or unsatisfactory which indicated that these ratings were sensitive enough to identify performance deficiencies among residents.

Key words : Inter-rater reliability, general surgery residents, clinical performance assessment

*Department of Surgery, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand.

เรื่องย่อ : ความถูกต้องและความแม่นยำของผลการประเมินการปฏิบัติงานของแพทย์ประจำบ้าน สาขาศัลยศาสตร์ทั่วไป

เชิดศักดิ์ ไอรมนิรัตน์ พ.บ., ว.ว. (ศัลยศาสตร์ทั่วไป)*

*ภาควิชาศัลยศาสตร์, คณะแพทยศาสตร์ศิริราชพยาบาล, มหาวิทยาลัยมหิดล, กรุงเทพมหานคร 10700.

สารบัญชารายชื่อ 2547; 56: 335-343.

ผลการประเมินการปฏิบัติงานของแพทย์ประจำบ้านในหลักสูตรศัลยศาสตร์ทั่วไป ภาควิชาศัลยศาสตร์ คณะแพทยศาสตร์ศิริราชพยาบาล ในช่วงปีการศึกษา 2543 - 2544 และ 2544 - 2545 ได้รับการวิเคราะห์เพื่อประเมินความถูกต้องและความแม่นยำของการประเมินผลโดยคณาจารย์ผู้ดูแลแพทย์ประจำบ้าน ความเห็นพ้องระหว่างอาจารย์ในการประเมินผลแพทย์ประจำบ้านได้รับการวิเคราะห์โดยการศึกษาค่าความผันแปรในกลุ่มคะแนนเปรียบเทียบระหว่างความผันแปรของคะแนนในแพทย์ประจำบ้านแต่ละคน กับความผันแปรของคะแนนของแพทย์ประจำบ้านทั้งกลุ่ม ความสัมพันธ์ระหว่างผลการประเมินทักษะแต่ละด้านของแพทย์ประจำบ้านได้รับการศึกษาเพื่อตรวจสอบความสามารถของอาจารย์ผู้ประเมินในการแยกแยะทักษะทางคลินิกของแพทย์ประจำบ้าน สัดส่วนของผลการประเมินที่ไม่เป็นที่น่าพอใจหรือกำกวมเมื่อเทียบกับผลการประเมินที่เป็นที่พอใจได้รับการศึกษาเพื่อตรวจสอบความไวของการประเมินในการชี้บ่งความบกพร่องของแพทย์ประจำบ้าน

จากข้อมูลของแพทย์ประจำบ้าน 42 คนในปีการศึกษา 2543 - 2544 พบว่าสัมประสิทธิ์ความเห็นพ้องระหว่างอาจารย์ผู้ประเมินอยู่ในช่วง 0.06 - 0.93 โดยมีค่าเฉลี่ย 0.51 จากข้อมูลของแพทย์ประจำบ้าน 47 คนในปีการศึกษา 2544 - 2545 พบว่าสัมประสิทธิ์ความเห็นพ้องระหว่างอาจารย์ผู้ประเมินอยู่ในช่วง 0.04 - 0.85 โดยมีค่าเฉลี่ย 0.49 สัมประสิทธิ์ความเห็นพ้องระหว่างอาจารย์ผู้ประเมินแพทย์ประจำบ้านชั้นปีที่ 3 อยู่ในเกณฑ์ยอมรับได้ แต่ผลการประเมินแพทย์ประจำบ้านชั้นปีที่ 1 และ 2 ยังไม่แม่นยำเท่าที่ควร ความไม่แม่นยำของการประเมินแพทย์ประจำบ้านชั้นปีที่ 1 และ 2 นี้ น่าจะเป็นผลจากการที่บทบาทในการปฏิบัติงานของแพทย์ประจำบ้านชั้นปีที่ 1 และ 2 มีน้อยและไม่เพียงพอที่อาจารย์จะประเมินได้อย่างแม่นยำ จากการศึกษาความสัมพันธ์ระหว่างผลการประเมินทักษะแต่ละด้านของแพทย์ประจำบ้านพบว่าทักษะทางคลินิกหลายด้านไม่ได้รับการแยกแยะโดยอาจารย์ผู้ทำการประเมิน ทักษะทางคลินิกเหล่านี้ได้แก่ มนุษยสัมพันธ์กับผู้ป่วยและญาติ มนุษยสัมพันธ์กับแพทย์ และมนุษยสัมพันธ์กับผู้ร่วมงานอื่นๆ ความรู้และการตัดสินใจทางคลินิก ความตั้งใจทำงานและสมรรถภาพในการปฏิบัติงาน, และ ความตรงต่อเวลา และความรับผิดชอบ ผลการประเมินที่ไม่เป็นที่น่าพอใจหรือกำกวมมีสัดส่วนประมาณร้อยละ 10 ของผลการประเมินทั้งหมด ซึ่งชี้บ่งว่าผลการประเมินเหล่านี้มีความไวเพียงพอในการตรวจสอบความบกพร่องของทักษะทางคลินิกของแพทย์ประจำบ้าน

INTRODUCTION

Evaluation of the performance of residents has received increased attention, not only from the medical community but also from the general population. This attention stems from several sources, including the speed of the development and dissemination of new medical information, data suggesting that medical knowledge declines over time, and concerns about the credentialing and licensure processes.^{1,2} Despite the significant increase of objective evaluations using in-training paper examinations and Objective Structured Clinical Examination (OSCE) in many residency programs, the predominant method of resident evaluation is still the subjective clinical performance evaluation by attending faculty.^{3,4}

Although the use of clinical performance evaluation forms has persisted in the medical communities for a long time, validity and reliability studies of their use in medical student and resident evaluation showed conflicting results. Some studies have shown that faculty ward evaluations of residents were unreliable, inflated, unable to discriminate between different clinical skills, and did not correlate with other measures of performance.⁴⁻⁶ However, many studies demonstrated faculty ward evaluation to be sufficiently reliable to provide precise estimates of the faculty's view of medical students and residents.^{3,7-9}

When performance assessments are being used to make decisions about individual residents or their residency programs, their validity and reliability must be held to a high standard. The higher the stakes for assessments, the more rigorous must be the evidence for the validity of the assessments.¹⁰ Reliability of scores is a necessary condition for the validity of inferences and decisions based on performance assessments, no less so than other kinds of assessment. Because resident performance assessment uses many faculty members to assign scores, the most important reliability evidence to be considered is inter-rater reliability.

The purpose of this study was to evaluate the inter-rater reliability of clinical performance assessment of residents in the general surgery residency program at the Department of Surgery, Faculty of Medicine Siriraj Hospital. Furthermore,

validity evidence of using the clinical performance rating in the evaluation of resident performance was also studied, specifically in the aspect of internal structure and educational discriminant validity. The results from this study will be useful for the Department of Surgery, Faculty of Medicine Siriraj Hospital, to support or refute the use of the clinical performance rating in resident evaluations. The results will also guide the direction to improve the quality of the clinical performance rating.

MATERIALS AND METHODS

This was a retrospective study of resident evaluations of all residents in the general surgery residency program of the Department of Surgery, Faculty of Medicine Siriraj Hospital, during the 2000 - 2001 and 2001 - 2002 academic years. This study included all three years of general surgery residents, the first three years of cardiothoracic residents and the first year of urology residents, pediatric surgery residents, neurosurgery residents, and orthopedic surgery residents.

Clinical performance of each resident was evaluated by attending faculty members at the end of each clinical rotation using a standard rating form containing 11 specific performance items: knowledge, clinical judgment, curiosity, work concentration, work effectiveness, medical records, punctuality, responsibility, and relationships with patients, other doctors, and other workers. Each performance rating was independently scored on the scale from one (very bad) to five (very good). The scores from all 11 items were summed up and converted into a percentage score of overall performance of each rotation. All residents in the same year of training were evaluated with the same evaluation standards, regardless of their specialties.

To evaluate the inter-rater reliability of clinical performance ratings, the intraclass correlation coefficients were calculated. This ANOVA-based type of correlation measures the relative homogeneity within groups in ratio to the total variation which is suitable for measuring inter-rater reliability when sample size is small or when there are more than two raters.¹¹

Table 1. Characteristics of studied population.

Year gender	2000 - 2001			2001 - 2002		
	First year	Second year	Third year	First year	Second year	Third year
Male	20 (86.96%)	10 (83.33%)	7 (100%)	22 (91.67%)	9 (81.82%)	10 (83.33%)
Female	3 (13.04%)	2 (16.67%)	0 (0%)	2 (8.33%)	2 (18.18%)	2 (16.67%)

Internal structure evidence of validity of the clinical performance rating was further explored with the Pearson correlation study between individual performance ratings. This demonstrated how well faculty members differentiated their ratings among individual performance skills.

Finally, educational discriminant validity was determined by the number of ratings that was in the marginal or unsatisfactory ranges, compared to satisfactory ratings. This study demonstrated the sensitivity of performance ratings to performance deficiencies of residents.

RESULTS

Resident characteristics

In the 2000 - 2001 academic year, 42 residents worked in the general surgery residency program. The number of residents in the first, second, and third years was 23, 12, and 7, respectively. In the first year, there were nine general surgery residents, four neurosurgery residents, four orthopedic surgery residents, three urology residents, two cardiothoracic residents, and one pediatric surgery resident. In the second year, there were ten general surgery residents and two cardiothoracic residents. All the residents in the third year were general surgery residents. Gender distribution of these residents can be summarized in Table 1.

In the 2001 - 2002 academic year, 47 residents worked in the program. There were 24 first-year, 11 second-year, and 12 third-year residents. In the first year, there were ten general surgery residents, four orthopedic surgery residents, three urology residents, three cardiothoracic residents, two neurosurgery residents, and two pediatric surgery residents. In the second year, there were nine general surgery residents

and two cardiothoracic residents. Ten of the third-year residents were general surgery residents and two were cardiothoracic residents. Gender distribution of these residents can be summarized in Table 1.

Rater characteristics

Raters of clinical performance evaluations of residents were attending faculty members of the services with whom that they worked. All residents in the same level of training had to rotate through the same services, but in different sequences. In the 2000-2001 academic year, all first-year residents worked in the services of pediatric surgery, urology, neurosurgery, plastic surgery, anesthesiology, intensive care unit, orthopedic surgery, trauma surgery, cardiothoracic surgery, head-neck-breast surgery, and two general surgery rotations. The second-year residents worked in the services of cardiothoracic surgery, urology, neurosurgery, pathology, the intensive care unit, two general surgery rotations, and two surgical rotations in provincial hospitals. The third-year residents worked in one pediatric surgery rotation, one head-neck-breast surgery rotation, one surgical rotation in a provincial hospital, three general surgery rotations, and three trauma surgery rotations.

In the 2001 - 2002 academic year, there was a revision in residency rotations. One rotation of minor operative surgery had been added into the first year residency which resulted in a total of 13 rotations. The second-year residents worked in 11 services including pathology, urology, cardiothoracic surgery, neurosurgery, plastic surgery, orthopedic surgery, gynecology, head-neck-breast surgery, trauma surgery, provincial hospital, and general surgery. Finally, one rotation in surgical division of a provincial hospital had been added to the third-year rotations.

Intraclass correlation

Considering there were two sources of score variability; subjects and raters, the intraclass correlation was studied with two-way ANOVA models. Due to a fixed set of raters for each group of residents, two-way mixed models were used. According to the way these rating scores were handled, the systematic differences among levels of ratings were considered relevant. The intraclass correlation was then measured with an absolute

agreement standard. Finally, in choosing the appropriate value of intraclass correlation coefficients, the average measure intraclass correlations were chosen because the decision about these residents were always made based on average ratings. Inter-rater reliability estimates for clinical performance evaluations based on the average measure intraclass correlations of two-way mixed models obtained from the analyses can be summarized in Tables 2 and 3.

Table 2. Intraclass correlation coefficients of clinical performance ratings of residents in the 2000 - 2001 academic year.

Residency year	1	2	3
Number of subjects	23	12	7
Knowledge	0.64	N/A*	0.93
Clinical judgment	0.26	0.25	0.84
Curiosity	0.55	0.63	0.79
Work concentration	0.52	0.67	0.69
Work effectiveness	0.57	0.19	0.73
Medical records	0.06	0.18	0.63
Punctuality	0.30	0.49	0.79
Responsibility	0.28	0.63	0.83
Relationship with patients	0.25	0.63	0.48
Relationship with other doctors	0.29	0.42	0.67
Relationship with other workers	0.21	0.46	0.55
Overall performance	0.53	0.58	0.89

* N/A = Not available

Table 3. Intraclass correlation coefficients of clinical performance ratings of residents in the 2001 - 2002 academic year.

Number of subjects	1	2	3
Residency year	1	2	3
Number of subjects	24	11	12
Knowledge	0.37	0.18	0.81
Clinical judgment	0.32	0.04	0.85
Curiosity	0.47	N/A*	0.75
Work concentration	0.31	N/A*	0.70
Work effectiveness	0.38	0.17	0.74
Medical records	0.26	N/A*	0.59
Punctuality	0.45	N/A*	0.69
Responsibility	0.29	N/A*	0.72
Relationship with patients	0.21	0.18	0.78
Relationship with other doctors	0.39	N/A*	0.77
Relationship with other workers	N/A*	N/A*	0.82
Overall performance	0.45	N/A*	0.83

*N/A = Not available

In the 2000 - 2001 academic year, reliability coefficients of individual ratings were in the range of 0.06 - 0.93 with an average value of 0.51. The inter-rater reliability of ratings in knowledge of second year residents could not be determined because variation among subjects was less than the variation within subjects, which indicated that there were nonrandom effects introduced by other variables.¹² The inter-rater reliabilities for overall performance of the first, second, and third year residents were 0.53, 0.58, and 0.89, respectively.

In the 2001 - 2002 academic year, reliability coefficients of individual ratings ranged from 0.04 to 0.85 with an average value of 0.49. Nine ratings had more variation within subjects than variation among subjects which resulted in undetermined reliability coefficients. Almost all of these undetermined results occurred in the ratings of second-year residents, including ratings of their overall performance. The inter-rater reliabilities of overall performance of the first and third-year residents were 0.45 and 0.83, respectively.

Intercorrelation between performance ratings

A Pearson correlation study between individual ratings was done. The result is summarized in Table 4. The correlation coefficients among items ranged from 0.47 to 0.94. Most of the performance rating items had strong relationships with one another. Fifty-one percent of these relationships demonstrated that one item had more than half of the score variability explainable by the score differences in another item. Six of these relationships between items (10.91%) were very strong to the degree that more than 75% of score differences in one item can be explained by variability of score in another item. These pairs of highly-correlated items are: knowledge and clinical judgment ($r = 0.88$), work concentration and work effectiveness ($r = 0.87$), punctuality and responsibility ($r = 0.88$), relationship with patients and relationship with other doctors ($r = 0.94$), relationship with patients and relationship with other workers ($r = 0.93$), and relationship with other doctors and relationship with other workers ($r = 0.94$).

Table 4. Intercorrelation matrix for resident performance ratings.

	I-1	I-2	I-3	I-4	I-5	I-6	I-7	I-8	I-9	I-10	I-11	OP
I-1	1	0.88*	0.74*	0.63*	0.71*	0.61*	0.61*	0.59*	0.47*	0.48*	0.50*	0.77*
I-2		1	0.77*	0.72*	0.75*	0.65*	0.67*	0.70*	0.55*	0.53*	0.56*	0.83*
I-3			1	0.84*	0.83*	0.67*	0.77*	0.80*	0.75*	0.75*	0.73*	0.92*
I-4				1	0.87*	0.66*	0.77*	0.86*	0.70*	0.70*	0.65*	0.89*
I-5					1	0.70*	0.73*	0.79*	0.66*	0.69*	0.68*	0.90*
I-6						1	0.72*	0.71*	0.64*	0.62*	0.65*	0.80*
I-7							1	0.88*	0.70*	0.71*	0.66*	0.87*
I-8								1	0.73*	0.74*	0.70*	0.90*
I-9									1	0.94*	0.93*	0.85*
I-10										1	0.94*	0.85*
I-11											1	0.84*
OP												1

I-1 = Item 1: Knowledge

I-2 = Item 2: Clinical judgment

I-3 = Item 3: Curiosity

I-4 = Item 4: Work concentration

I-5 = Item 5: Work effectiveness

I-6 = Item 6: Medical records

I-7 = Item 7: Punctuality

I-8 = Item 8: Responsibility

I-9 = Item 9: Relationship with patients

I-10 = Item 10: Relationship with other doctors

I-11 = Item 11: Relationship with other workers

OP = Overall Performance

* Correlation is significant at the 0.01 level (two-tailed)

Educational discriminant validity

From 4,740 ratings in the 2000 - 2001 academic year, 491 (10.36%) were marginal (rating of 3), and eight (0.17%) were unsatisfactory (rating below 3). In the 2001 - 2002 academic year, there were 6,044 ratings. Five hundred and twenty-eight (8.74%) of them were marginal. Fifty-five (0.91%) of them were unsatisfactory.

DISCUSSION

Faculty evaluations of clinical performance of residents have become more increasingly important as the clinical competence of residents has come under closer scrutiny. Validity evidence of the use of clinical performance assessments is needed to ensure the integrity of these instruments in making decisions about individual residents as well as the residency programs. This study provides some validity evidence for the use of clinical performance rating of residents in the general surgery residency program of the Department of Surgery, Faculty of Medicine Siriraj Hospital.

From intraclass correlation study, the reliability coefficients for ratings of first and second-year residents were lower than those of third-year residents. In the 2000 - 2001 academic year, reliability coefficients of performance ratings of first and second-year residents were in the range of 0.06 - 0.67 with an average of 0.40, while those of third-year residents were in the range of 0.48 - 0.93 with an average of 0.72. This pattern was consistent with findings from the 2001 - 2002 academic year, in which reliability coefficients of ratings of first and second-year residents were in the range of 0.04 - 0.47 with an average of 0.29, while those of third-year residents were in the range of 0.59 - 0.85 with an average of 0.75.

Considering the stakes associated with these clinical performance ratings, the inter-rater reliability estimates of performance ratings of third-year residents were at an acceptable standard, while those of the first and second-year residents need improvement. These performance ratings can be considered to be medium-stake assessments like those of classroom achievement tests which usually require a reliability coefficient in the range of 0.65 - 0.70. It should also be noted that ratings that had variations within subject exceeded variations between subjects,

which resulted in undetermined reliability coefficients, were all ratings of the first and second-year residents.

The possible explanations for the more reliable clinical performance ratings of third-year residents could be from two sources of score variability: raters and subjects. In order to check whether raters of third-year residents had higher inter-rater reliability than raters of first and second-year residents, additional intraclass correlation analyses were done for first and second-year residents using only data from raters in services that evaluated third-year residents: general surgery, trauma surgery, head-neck-breast surgery, and pediatric surgery. In the 2000 - 2001 academic year, the reliability coefficients were in the range of 0.11 - 0.82 with an average of 0.41. In the 2001 - 2002 academic year, the reliability coefficients ranged from 0.11 to 0.45 with an average of 0.25. Reliability coefficients did not seem to improve when limiting the analysis to only these rotations. Thus, the difference in raters between years could not explain the finding.

The other explanation would be the differences in the roles of residents to be evaluated. One of the validity challenges faced by performance assessment is the inadequate task to be evaluated. Large numbers of tasks are needed to get a reliable assessment.¹³ With the training system used at the Department of Surgery, Faculty of Medicine Siriraj Hospital, first and second-year residents had very limited roles in every service. Almost all the responsibility was put in the hands of third-year residents. Tasks performed by first and second-year residents were inadequate for faculty members to reliably evaluate. The Department of Surgery should consider giving first and second-year residents more responsibility in patient care service in order to have adequate task sampling for reliable clinical performance evaluation.

Another consideration of inter-rater reliability is rater leniency. The absolute agreement measures used in the above analyses took into account both variability from rater leniency and variability from resident performance. To check whether rater leniency had significant effects on inter-rater reliability, further intraclass correlation analysis was performed using consistency measures. The consistency measures considered variability due to

levels of ratings irrelevant. What mattered was the relative standing of residents, not the absolute value of scores assigned by raters. It was found that with consistency measure, inter-rater reliability could be increased by 11.95% from using absolute agreement measures. Rater leniency or stringency seemed to have some effect on inter-rater reliability. However, the main issue seemed to be the difference in the criteria used to judge clinical performance of residents.

From the correlation study between individual ratings, most of the performance rating items had a strong relationship with one another. This suggested that faculty members did not discriminate among different clinical skills very well in evaluating these residents. Many performance rating items seemed to be interpreted by raters to be the same thing. The most significant ones were relationship with patients, relationship with other doctors, and relationship with other workers. Faculty members tended to make one overall, undifferentiated judgment on these three items. Other noteworthy relationships were relationships between knowledge and clinical judgment, work concentration and work effectiveness, and punctuality and responsibility. Faculty members need some guidelines to differentiate among these items in evaluating resident performance.

Analysis of educational discriminant validity showed that 10 percent of ratings were marginal or unsatisfactory. Compared to other studies which had only 1.3% of ratings for marginal or unsatisfactory performance, performance ratings of residents at the Department of Surgery, Faculty of Medicine Siriraj Hospital, worked well in detecting performance deficiencies among residents.³

The results from this study can be used to guide the direction to improve the quality of clinical performance evaluation of residents at the Department of Surgery, Faculty of Medicine Siriraj Hospital. First and second-year residents should have more active roles in every service to provide adequate tasks to be evaluated. Although inter-rater reliabilities of performance ratings for third-year residents are acceptable for medium-stakes assessment, as in the monitoring of training progression, they also need improvement if a major decision has to be made from these assessments such as recommendation for board examinations or termination of training. A high-stake

decision usually requires reliability above 0.90. Evidence from literature shows that high rater reliability is quite possible and feasible to achieve if there are specific scoring guidelines and sufficient training for raters.^{10,13,14} The Department of Surgery, Faculty of Medicine Siriraj Hospital, should develop scoring guidelines for rating clinical performance of residents, not only to improve inter-rater reliability of scores but also to improve the differentiation among performance items in the rating form.

From the analysis of educational discriminant validity, the clinical performance ratings of the Department of Surgery, Faculty of Medicine Siriraj Hospital, were sensitive to performance deficiencies among residents. Faculty members should use information obtained from these ratings and give feedback to individual residents to help them improve their performance.

Another issue in clinical performance assessment of residents is the limited number of observations of attending faculty on residents. Alternative raters (including peers, self, nurses, and patients) have been used in some studies. These alternative raters can provide information complementary to faculty ratings that is useful in the overall performance assessment of residents.^{7,15} The Department of Surgery, Faculty of Medicine Siriraj Hospital, might consider using these alternative raters as complementary sources of information.

CONCLUSION

Inter-rater reliability, internal structure, and educational discriminant validity of the clinical performance assessment of residents in the general surgery residency program at the Department of Surgery, Faculty of Medicine Siriraj Hospital, have been studied. Inter-rater reliabilities of performance ratings of third-year residents were at an acceptable standard for such a medium-stake assessment. However, ratings of first and second-year residents need some improvement. Intercorrelation among performance ratings demonstrated the lack of discrimination between some clinical skills in faculty ratings. Finally, the clinical performance ratings seem to be sensitive enough to identify performance deficiencies among residents.

REFERENCES

1. Ramsey PG, Carline JD, Inui TS, et al. Changes over time in the knowledge base of practicing internists. JAMA 1991; 266: 1103-7.
2. Ramsey PG, Wenrich MD, Carline JD, et al. Use of peer ratings to evaluate physician performance. JAMA 1993; 269: 1655-60.
3. Kwolek CJ, Donnelly MB, Sloan DA, et al. Ward evaluations: Should they be abandoned? J Surg Res 1997; 69: 1-6.
4. Sloan DA, Donnelly MB, Drake D, Schwartz RW. Faculty sensitivity in detecting medical students' clinical competence. Med Teach 1995; 95: 335-42.
5. Schwartz RW, Donnelly MB, Sloan DA, Johnson SB, Strodel WE. Assessing senior residents' knowledge and performance: An integrated evaluation program. Surgery 1994; 116: 634-37.
6. Maxim BR, Dielman TE. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. Med Educ 1987; 21: 130-37.
7. Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol 2002; 99: 647-51.
8. Keyman A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. Med Educ 1987; 21: 477-81.
9. 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: 506-10.
10. Moore AD, Young S. Clarifying the blurred image: Estimating the inter-rater reliability of performance assessments. Annual meeting of the Northern Rocky Mt. Educational Research Association. Jackson, Wyoming, 1997: 1-19.
11. Garson GD. Reliability. In: Quantitative research in public administration. Spring 2004. Available at <http://www2chass.ncsu.edu/garson/pa765/reliab.htm> Accessed Jan 2004.
12. Garson GD. Correlation. In: Quantitative research in public administration. Spring 2004. Available at <http://www2chass.ncsu.edu/garson/pa765/correl.htm> Accessed Jan 2004.
13. Shavelson RJ, Gao X, Baxter GP. Sampling variability of performance assessments. Report on the Status of Generalizability Performance: Generalizability and transfer of performance assessments. Project 2.4: Design theory and psychometrics for complex performance assessment in science. National Center for Research on Evaluation, Standards, and Student Testing 1993.
14. Miller-Whitehead M. Inter-rater reliability on performance criteria: Theoretical issues. Tennessee Valley Educators for Excellence, 2001.
15. Johnson D, Cujec B. Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. Crit Care Med 1998; 26: 1811-16.