นิพนธ์ต้นฉบับ

Validity and reliability of 5, 7 and 9-point rating scales in Medical Education[*]

Boonnart Laisnitsarekul[**]
Sirichai Kanjanawasee[***]

Laisnitsarekul B, Kanjanawasee S. Validity and reliability of 5, 7 and 9-point rating scales in Medical Education. Chula Med J 1993 Jun; 37(6) : 387-395

The objectives of this basic research project were to find the validity and reliability of the 5, 7 and 9- point rating scales and compare the validity and reliability among them in clinical performance evaluation. Twenty-seven evaluators from Bhumibol Adulyadej Hospital, the Royal Thai Air Force, used the 5, 7 and 9- point rating scales with checklists to evaluate 39 medical students in nine stations of the Objective Structured Clinical Examination (OSCE) : OB-GYN history taking, wound suturing, knotting, routine history taking, shifting dullness, opthalmoscopy, tepid sponge, deep tendon reflex, and artificial milk feeding.

The concurrent validity of the 7-point rating scale was significantly higher (P<.05) than the 9-point rating scale in two stations : wound suturing and knotting. Also, the concurrent validity of the 7-point rating scale was significantly higher (P<.05) than the 5-point rating scale in one station opthalmoscopy. The intra-rater reliability for the nine stations were between 0.44 and 0.91 . The internal consistency reliability of the 5, 7 and 9-point rating scales were 0.12, 0.20 and 0.29, respectively. When compared among them, there was no difference in the internal consistency reliability.

Key words : *Validity, Reliability, Rating scale, OSCE, Medical Education.*

Reprint request : Laisnitsarekul B, Medical Education Unit Faculty of Medicine,Chulalongkorn University, Bangkok 10330, Thailand.

Received for publication. January 25, 1993.

[*] The Rachadapiseksompoj-China Medical Board Research Funds.

[**] Medical Education Unit, Faculty of Medicine, Chulalongkorn University.

[***] Department of Educational Research, Faculty of Education, Chulalongkorn University.

บุญนาท ลายสนิทเสรีกุล, ศิริชัย กาญจนวาสี. ความตรงและความเที่ยงของแบบมาตราส่วนประมาณค่าชนิด 5,7 และ 9 มาตราในการประเมินทางการศึกษาแพทยศาสตร์. จุฬาลงกรณ์เวชสาร 2536 มิถุนายน; 37(6) : 387-395

การวิจัยนี้เป็นการวิจัยพื้นฐานมีวัตถุประสงค์ เพื่อศึกษาความตรงและความเที่ยงของแบบมาตราส่วนประมาณค่าชนิด 5,7 และ 9 มาตรา และเปรียบเทียบค่าความตรงและความเที่ยง ระหว่างแบบมาตราส่วนประมาณค่าทั้งสามมาตรา ในการประเมินพฤติกรรมคลินิกทางการแพทย์

กลุ่มตัวอย่างที่ศึกษา ได้แก่ แพทย์โรงพยาบาลภูมิพลอดุลยเดช กรมแพทย์ทหารอากาศ จำนวน 27 คน เครื่องมือที่ใช้เป็นแบบมาตราส่วนประมาณค่าชนิด 5, 7 และ 9 มาตรา ร่วมกับแบบเลือกตรวจสำหรับประเมินพฤติกรรมคลินิก 9 ชนิดได้แก่ ทักษะในการซักประวัติทางสูติศาสตร์-นรีเวชวิทยา การเย็บแผล การผูกไหมเย็บ การซักประวัติผู้ป่วยที่มาด้วยอาการรู้สึกเพลียไม่มีแรง การตรวจหาสารน้ำในช่องท้องโดยวิธี shifting dullness การตรวจตาขวาโดยเครื่อง opthalmoscope การเช็ดตัวเด็กเพื่อลดไข้ การตรวจ deep tendon reflex และการให้คำแนะนำในการเลี้ยงลูกด้วยนมผง เครื่องมือดังกล่าวได้นำไปใช้ประเมินนิสิตแพทย์ชั้นปีที่ 5, 6 และแพทย์ฝึกหัด จำนวนรวม 39 คน ที่กำลังปฏิบัติงานอยู่ ณ โรงพยาบาลภูมิพลอดุลยเดช ในเดือนมีนาคม 2535

ผลการวิจัยที่สำคัญมีดังนี้

1. แบบมาตราส่วนประมาณค่าชนิด 7 มาตรา มีค่าความตรงร่วมสมัยสูงกว่าแบบมาตราส่วนประมาณค่าชนิด 9 มาตราด้านการประเมินเรื่องการเย็บแผล และการผูกไหมเย็บ นอกจากนี้แบบมาตราส่วนประมาณค่าชนิด 7 มาตรา มีค่าความตรงร่วมสมัยสูงกว่าแบบมาตราส่วนประมาณค่าชนิด 5 มาตรา ด้านการประเมินเรื่องการตรวจตาขวาโดยเครื่อง opthalmoscope

2. จากการตรวจสอบความเที่ยงภายในของผู้ประเมินมีค่าอยู่ระหว่าง 0.44 ถึง 0.91 และความเที่ยงแบบความสอดคล้องภายในของแบบมาตราส่วนประมาณค่าชนิด 5,7, และ 9 มาตรา มีค่าเท่ากับ 0.12, 0.20, และ 0.29 ตามลำดับ เมื่อเปรียบเทียบความเที่ยงแบบความสอดคล้องภายในไม่พบความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

A rating scale is a device by which a rater can record his judgement of another person (or of himself) based on the traits defined by the scale.[1] A rating scale consists of the name or description of the attribute to be measured, and a way of indicating the amount or quality of that attribute. The format of the scale can vary quite widely. In the simplest form, there would be two boxes, labelled for example, "satisfactory" and "unsatisfactory". To evaluate varying degrees of the trait, more boxes can be used (five to nine being the optimum in most situations), or the evaluator can place an "X" along a straight line. Ideally, each box should also have a description of the observed behaviours associated with the particular rating. Rating scales are most frequently used to evaluate people in areas where objective outcomes are not available or are not feasible. This includes areas such as "problem-solving ability", "initivative", and "self-directed learning".[2]

In constructing a rating scale, one of the popular questions is how many points there should be on the continuum. This could range from just two or three (Satisfactory-Unsatisfactory; Below Average-Average-Above Average) to simply labelling the two extremes and placing 20 boxes between them. Among the various scales which have been reported, the preference seems to be to use four or five points (e.g., Dielman et al., 1979, 1980;[3,4] Gough et al., 1964;[5] Linn, 1979;[6] ERIC CD-ROM, 1983-1990,[7] although as few as three (Cowles and Kubany, 1959;[8] Geertsma and Chapman, 1967)[9] and as many as 20 (Brumback and Howell, 1972)[10] have been employed.

Are five or seven or nine points sufficient to accurately differentiate among students? For many years, test construction theorists have debated the optimum number of points for a scale. If there are too few divisions, then the test may not be fully utilizing the rater's ability to discriminate fine gradations in performance from one person to another. Too many points, though, may be beyond the rater's powers of discrimination. Symonds[11] concluded that, on theoretical grounds, the optimum number of steps was seven.

Guilford[12] stated : "It can be said that the number 7 is usually lower than optimal and it may pay in some favorable situations to use up to 25 scale divisions". Nunnally,[13] also on theoretical grounds, stated that the reliability of a scale increased rapidly as the number of divisions increases to about seven, and then rises more slowly until there are 11 points. Bendig[14,15] found that a drop in reliability occurred when there were fewer than 4 or more than 10 categories. Thus, while the actual optimum number of steps is still a matter of some conjecture, it is probably between 7 and 11 more than the four or five divisions commonly used. It may seem that arguing for at least seven points to be used rather than five is quibbling over a minor matter. In Thailand, the five-point rating scale is commonly used. A review of the literature from JDEX 1962-1989[16] and Thai journals between 1990 and 1992 shown that no paper studied the points of a rating scale. This stimulated the authors' interest in studying the validity and reliability of the 5, 7 and 9-point rating scale in Medical Education.

Objectives

1. To find the concurrent validity, intra-rater reliability and internal consistency reliability of the 5, 7 and 9-point rating scale.

2. To compare the concurrent validity and internal consistency reliability among them in clinical performance evaluation.

Hypothesis

On theoretical grounds, Symonds (1924) concluded that the optimum number for a rating scale was seven. Bendig (1953,1954) found that there was a drop in reliability when there were fewer than four or more than 10 categories. Nunnally (1967) also stated that the reliability of a scale increased rapidly as the number of divisions increases to about seven, and then rises more slowly until there are 11 points. Therefore, the authors would like to adopt as their hypothesis the following :

1. Rating scales which have a different number of points should have differences in validity and reliability.

2. The validity and reliability of a seven-point scale should be higher than five or nine-point rating scales.

Materials

1. Population

1.1 Twenty-seven medical teachers from Bhumibol Adulyadej Hospital, the Royal Thai Air Force, were chosen as evaluators. There were three evaluators from the Department of Obstetrics and Gynecology, six evaluators from the Department of Surgery, nine evaluators from the Department of Medicine and nine evaluators from the Department of Pediatrics.

1.2 Thirty-nine medical students from the Faculty of medicine, Chulalongkorn University and Bhumibol Adulyadej Hospital, the Royal Thai Air Force, were chosen as examinees. There were 12 fifth-year students, 15 sixth-year students and 12 interns.

2. Tools

2.1 Checklists were planned to assess clinical skills at nine stations : OB-GYN history-taking, wound suturing, knotting, routine history-taking, shifting dullness, opthalmoscopy, tepid sponge, deep tendon reflex and artificial milk feeding. All of them were borrowed from the department and the clinical experience committee, Faculty of Medicine, Chulalongkorn University. Checklists contain a list of items (generally history questions/findings, examination maneuvers, and/or information to be communicated to the patient, depending upon the type of station), and evaluator's task is simply to check off the items done by the examinee and give a score. An example of a checklist from a deep tendon reflex station is shown in Figure 1.

2.2 Ratings typically use Likert-type scales, in which evaluators directly score examinee performance. A rating scale consists of items reflecting broader categories of performance. The five, seven and nine boxes are designed to assess the student's overall performance at the station, as previously dercribed. Each may not be anchored with specific behavioral descriptions. Figure 1 is an example of a seven point scale from a deep tendon reflex station.

ITEM	Weight	Score
1. Ask for permission	1	
2. Explain the steps of physical examination	2	
3. Check knee reflex	2	
4. Check ankle reflex	2	
5. Check Babinski reflex	2	
6. Thank the patient	1	
TOTAL SCORE	10	

The competence of this student is

highest	7	6	5	4	3	2	1	lowest

Figure 1. Checklist and rating scale of deep tendon reflex station.

Method

The examination will be called an Objective Structured Clinical Examination (OSCE) and will be aimed at testing nine clinical skills including : history taking, physical examination, communication and procedural skills. (Figure 2.) The OSCE will be four hours long; all stations were five minutes in length. The 27 evaluatiors from Bhumibol Adulyadej Hospital, the Royal Thai Air Force, used 5, 7 and 9-point rating scales with checklists to evaluate 39 medical students. In Figure 3, a student enters the examination at station 1 and the three senior evaluators used the same checklist but a different rating scale to rate the student's performance according to or her prospectus. When the second student enters station 1, each evaluator will use a different rating scale by systematic random sampling. For analysing data, the researcher calculated inter-rater agreement of the checklists for individual stations by Hoyt's Analysis of Variance(17) and used it as the external criteria. The concurrent validity of the 5, 7 and 9-point rating scales would correlate them with the average score of the checklist and compare among them by inferential technique.(18) Two kinds of reliability were calculated. Intra-rater reliability, as measured by Pearson product moment correlation between the score of the measured by Hoyt's Analysis of Variance, indicates whether all of the stations are measuring the same domain. Inferential technique was used to compare internal consistency reliability among the 5, 7 and 9-point rating scales.

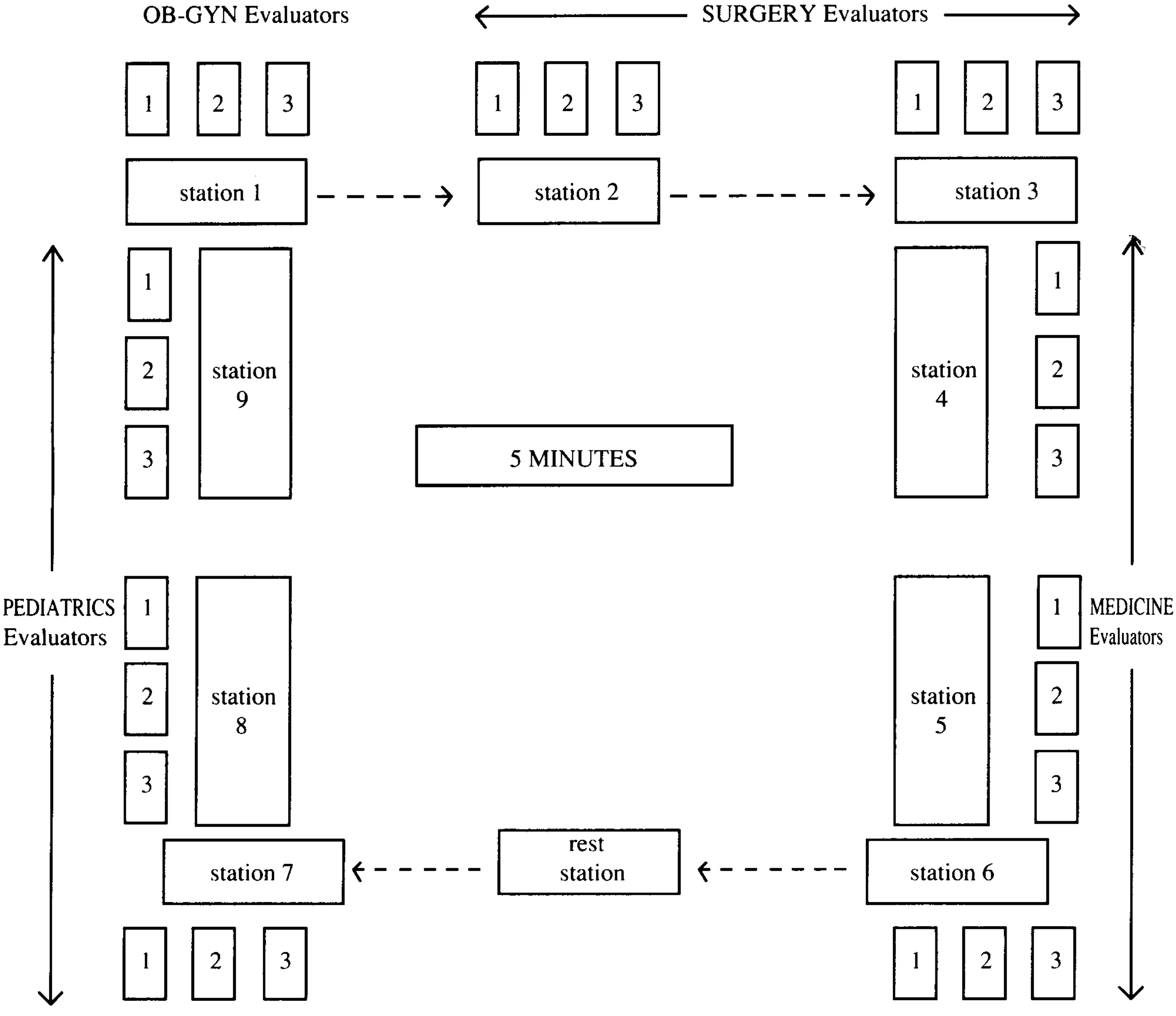
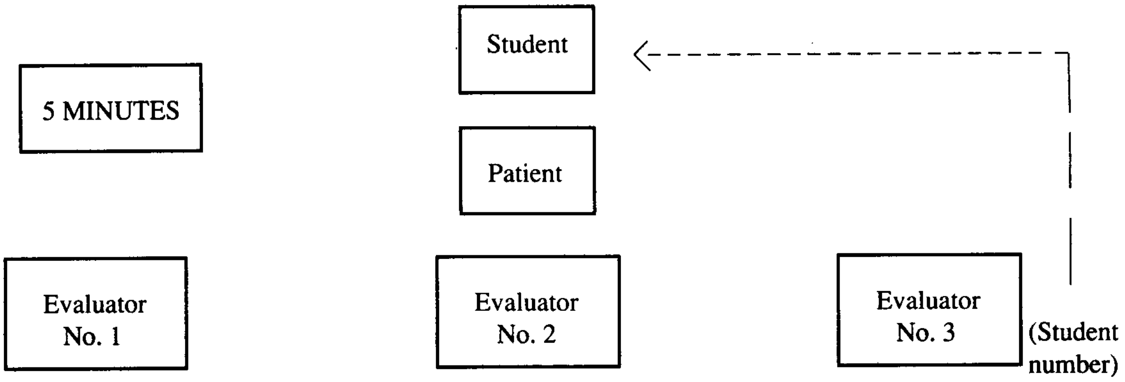


Figure 2. Complete structured clinical examination using nine stations.



C		R	5	7	9	1
H		A	7	9	5	2
E		T	9	5	7	3
C		I	5	7	9	4
K	+	N	7	9	7	5
		G	9	5	5	6
L			.	.	.	.
I		S	.	.	.	.
S		C	.	.	.	.
T		A	.	.	.	.
		L	.	.	.	.
		E	.	.	.	39

Figure 3. An example of data collection using checklist and rating scale.

Results

The inter-rater agreement of checklists in nine stations was as follow : 0.90, 0.74, 0.65, 0.91, 0.86, 0.89, 0.68, 0.53 and 0.80 (Table 1). In the nine stations, the concurrent validity of the 5, 7, 9-point rating scales was between 0.47 and 0.78 (5-point), between 0.54 and 0.84 (7-point), between 0.23 and 0.83 (9-point), respectively. The concurrent validity of the 7-point was significantly higher than the 9-point in two stations : Wound suturing and Knotting (P<.05). Also the concurrent validity of the 7-point rating scale was significantly higher than the 5-point in one station : Opthalmoscopy (P<.05) (Table 2). The intra-rater reliability, correlation between checklist scores and ratings in the nine stations were between 0.44 and0.91(5point) between 0.56 and 0.87 (7-point scale), between 0.68 and 0.90(9-point), respectively (Table 2). The internal consistency reliability levels of the 5, 7, 9-point rating scales were 0.12, 0.20, and 0.29, respectively. When compared among them, there was as no difference in internal consistency reliability.

Table 1. The inter-rater agreement of checklists in nine stations.

Station	Reliability
1. OB-GYN : Taking history	0.90**
2. SURGERY : Suturing	0.74**
3. SURGERY : Knotting	0.65**
4. MEDICINE : Taking history	0.91**
5. MEDICINE : Shifting dullness	0.86**
6. MEDICINE : Opthalmoscopy	0.89**
7. PEDIATRICS : Tepid sponge	0.68**
8. PEDIATRICS : Deep tendon reflex	0.53**
9. PEDIATRICS : Feeding	0.80**

** P < .01

Table 2. Concurrent validity and intra-rater reliability of the 5, 7 and 9-point rating scales in nine stations.

Station	Rating scale	Validity	Reliability
	(points)		
1. OB-GYN : Taking history	5	0.75**	0.72**
	7	0.70**	0.72**
	9	0.70**	0.72**
2. SURGERY : Suturing	5	0.53**	0.63**
	7	0.64**	0.56**
	9	0.23**	0.68**
3. SURGERY : Knotting	5	0.71**	0.91**
	7	0.82**	0.87**
	9	0.63**	0.90**
4. MEDICINE : Taking history	5	0.66**	0.76**
	7	0.68**	0.76**
	9	0.79**	0.81**
5. MEDICIDNE : Shifting dullness	5	0.74**	0.82**
	7	0.76**	0.84**
	9	0.76**	0.89**
6. MEDICINE : Opthalmoscopy	5	0.47**	0.44**
	7	0.76**	0.75**
	9	0.65**	0.70**
7. PEDIATRICS : Tepid sponge	5	0.78**	0.64**
	7	0.84**	0.82**
	9	0.83**	0.75**
8. PEDIATRICS : Deep tendon reflex	5	0.60**	0.85**
	7	0.54**	0.82**
	9	0.47**	0.71**
9. PEDIATRICS : Feeding	5	0.76**	0.85**
	7	0.78**	0.87 **
	9	0.72**	0.88**

** P < .01

Discussion

The inter-rater agreement of checklists in the nine stations ranged from 0.44 to 0.91 (P<.01). This finding suggests that inter-rater agreement is fairly good for individual stations, and rater-to-rater variability is not a major source of measurement error for total test scores. When we correlated the score of the rating scales with the average score of the checklists, accepted criteria, measured at the same time, the concurrent validity of 5, 7 and 9-point rating scales are fairly good for individual stations. When comparing concurrent validity among them, the concurrent validity of the 7-point was significantly higer than the 5-point or 9- point(P<.05) in some stations. In addition, the correlation between checklist scores and the 5, 7, 9-point ratings averaged 0.74, 0.78, 0.78, respectively, suggesting that similar information was obtained with the two formats. Rating scales were completed after the checklists by the same rater; this may well have influenced the results. These findings suggest that rating scales are quite usable, particularly for areas where checklists are more difficult to develop (e.g. those to rater attitudes and aspects of communication skills) without trivializing the skill to be measured. From an educational perspective, checklists have the advantage of providing students with a better definition of expectations. However, this can lead students to memorize checklists without achieving a deeper understanding of the skill to be performed,[19] ratings may prevent this from occurring. More research in this area would be desirable; for the present, it seems appropriated to base the choice of checklists and rating forms on the skill to be measured. Inspection of Table 2 indicates that the concurrent validity and intra-rater reliability of the 7-point rating scale should be better than the 5-point or 9-point. However, since this research effort is the first of its type to be reported in Thailand, more research in this area would be desirable.

On the other hand, the internal consistency reliability of the 5, 7 and 9-point rating scales are low. In most settings, students are rated at multiple stations such as history-taking, physical examination and communication skills in different areas (OB-GYN, Surgery, Medicine, Pediatrics). Harden[20] suggested the OSCE not only be used as a summative form of evaluation but also formative form because it can provide feedback on progress to both staff and students. The results of the OSCE should provide a profile of a student. They can present a picture of the student's strengths and weakness in different areas such as history-taking, physical examination, problem solving and attitudes.

Summary

The objectives of this basic research project were to find the concurrent validity, intra-rater reliability and internal consistency reliability of the 5, 7 and 9-point rating scales and compare the concurrent validity and internal consistency reliability among them in clinical performance evaluation. The 27 evaluators from Bhumibol Adulyadej Hospital, the Royal Thai Air Force, used 5, 7 and 9-point rating scales with checklists to evaluate 39 medical students in nine stations of the Objective Structured Clinical Examination (OSCE), such as OB-GYN history taking, wound suturing, knotting, routine history taking, shifting dullness, opthalmoscopy, tepid sponge, deep tendon reflex, and artificial milk feeding. The inter-rater agreement levels of the checklists in nine stations were between 0.53 and 0.91 (P<.01). The concurrent validity of the 5, 7 and 9-point rating scales are fairly good for individual stations. When comparing the concurrent validity among them the concurrent validity of the 7-point was significantly higher than the 5-point or 9-point (P<.05) in some stations. In addition, the intra-rater reliability of the 5, 7 and 9-point rating scales averaged 0.74, 0.78 and 0.78, respectively. These findings suggest that all rating scales are quite usable, but the 7-point should give better resalts than the others. To measure the internal consistency of the rating scales by calculation in all stations. The OSCE should have a low level of agreement.

Acknowledgements

Without the support financial of the Rachadapiseksompoj-China Medical Board Research

Funds, this research project would have been extremely difficult, if not impossible, to conduct. The researchers are grateful to all the staff of the Department of Obstetrics and Gynecology, Surgery, Medicine and Pediatrics, Faculty of Medicine, Chulalongkorn University and Bhumibol Adulyadej Hospital, the Royal Thai Air Force, for their contribution towards the tools for and implementation of the project. We are indebted to the Director of Bhumibol Adulyadej Hospital, Group Capt. Dr. Pitoon Chuangpanich M.D. and Flight Lt. Dr. Jaitip piboon M.D., the Royal Thai Air Force, who helped us with data collection. We also would like to thank the officers of the Innovative Medical Unit, Faculty of Medicine, Chulalongkorn University for their contributions toward collecting the necessary equipment. We also indebted to Professor Dr. Chaloem Varavithya, Head of Research and Development for the Medical Education Centre, Faculty of Medicine, Chulalongkorn University who suggested the project and gave continuous advice.

References

1. Hilgard ER. Introduction to Psychology. 3rd ed. New York : Harcourt, Brace & World, 1962: 463
2. Dielman TE, Hull AL, Davis WK. Inter-rater Reliability of Clinical Performance Ratings. Proceedings, 18th annual conference on Research in Medical Education. Washington, D.C, 1979.
3. Streiner DL. Global rating scales. In : Neufeld VR, Norman GR, eds. Assessing Clinical Competence. New York : Springer, 1985: 119-41
4. Gough HG, Hall WB, Harris RE. Admission procedures as forecasters of performance in medical training. J Med Educ 1963 Dec; 38(12) : 983-98
5. Linn L. Interns' attitudes and values as antecedents of clinical performance. J Med Educ 1979 Mar; 54(3) : 238-40
6. Chulalongkorn University. ERIC CD-ROM 1983-1990. Bangkok : Academic Service Institute, Chulalongkorn University. 1991.
7. Cowles JT, Kubany AJ. Improving the measurement of clinical performance in medical students. J Clin Psychol 1959; 15 : 139-42
8. Geertsma RH, Chapman JE. The evaluation of medical students. J Med Educ 1967 Oct; 42(10) : 938-48
9. Brumback GB, Howell MA. Rating the clinical effectiveness of employed physicians. J Appl Psychol 1972 Jun; 56 : 241-4
10. Symonds PM. On the loss of reliability in ratings due to coarseness of the scale. J Exp Psychol 1924; 7 : 456-61
11. Guilford JP. Psychometric Methods. 2nd ed. New York : McGraw-Hill. 1954.
12. Nunnally JC. Psychometric Theory. New York : McGraw-Hill. 1967.
13. Bendig AW. The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. J Appl Psychol 1953; 37(1) : 38-41
14. Bendig AW. Reliability and numbers of rating scale categories. J Appl Psychol 1954; 38(1) : 38-40
15. Chulalongkorn University. JDEX 1962-1989. Bangkok : Faculty of Medicine Library, Chulalongkorn University. 1991.
16. Mehrens WA, Lehmann IJ. Standardized Tests in Education. 2nd ed. New York : Holt, Rinehart and Winston, 1975: 47
17. Glass VG, Stanley JC. Statistical Methods in Education and Psychology. New Jersey : Prentice-Hall, 1970: 313-4
18. Van Luijk S, Van der Vleuten C, Van Schelven R. The Relation between Content and Psychometric Characteri stics in Performance-based Testing. In : Bender W, Hiemstra R, Scherpbier A, Zwierstra R, eds. Teaching and Assessing Clinical Competence. Groningen : Boekwerk Publications. 1990.
19. Harden RM. The OSCE-A 15 Year Retrospective. In : Hart IR, Harden RM, Marchais JD, eds. Current Developments in Assessing Clinical Competence. Montreal : Can-Heal Publications, 1992: 41-53