Yeungnam University Journal of Medicine
YUJM

Original article
eISSN 2384-0293
Yeungnam Univ J Med 2021;38(2):118-126
https://doi.org/10.12701/yujm.2020.00423

A study on evaluator factors affecting physician-patient interaction scores in clinical performance examinations: a single medical school experience

Young Soon Park[1], Kyung Hee Chun[1], Kyeong Soo Lee[2], Young Hwan Lee[3]

[1]Department of Medical Education, Konyang University College of Medicine, Daejeon, Korea
[2]Department of Preventive Medicine and Public Health, Yeungnam University College of Medicine, Daegu, Korea
[3]Department of Medical Humanities, Yeungnam University College of Medicine, Daegu, Korea

Received: June 2, 2020
Revised: June 25, 2020
Accepted: June 25, 2020

Corresponding author:
Young Hwan Lee, MD, PhD
Department of Medical Humanities, Yeungnam University College of Medicine, 170 Hyunchoong-ro, Nam-gu, Daegu 42415, Korea
Tel: +82-53-640-6999
Fax: +82-53-620-2252
E-mail: yhlee3535@ynu.ac.kr

Background: This study is an analysis of evaluator factors affecting physician-patient interaction (PPI) scores in clinical performance examination (CPX). The purpose of this study was to investigate possible ways to increase the reliability of the CPX evaluation.
Methods: The six-item Yeungnam University Scale (YUS), four-item analytic global rating scale (AGRS), and one-item holistic rating scale (HRS) were used to evaluate student performance in PPI. A total of 72 fourth-year students from Yeungnam University College of Medicine in Korea participated in the evaluation with 32 faculty and 16 standardized patient (SP) raters. The study then examined the differences in scores between types of scale, raters (SP vs. faculty), faculty specialty, evaluation experience, and level of fatigue as time passes.
Results: There were significant differences between faculty and SP scores in all three scales and a significant correlation among raters' scores. Scores given by raters on items related to their specialty were lower than those given by raters on items out of their specialty. On the YUS and AGRS, there were significant differences based on the faculty's evaluation experience; scores by raters who had three to ten previous evaluation experiences were lower than others' scores. There were also significant differences among SP raters on all scales. The correlation between the YUS and AGRS/HRS declined significantly according to the length of evaluation time.
Conclusion: In CPX, PPI score reliability was found to be significantly affected by the evaluator factors as well as the type of scale.

Keywords: Clinical competence; Medical students; Physician-patient relations

Introduction

Establishing major competencies and training for clinical performance in medical education has been prevalent in Korea for only the last 15 years. This is due to the clinical skills test that was added to the paper-based Korean Medical Licensing Examination (KMLE) in 2009. This crucial change in the KMLE has strengthened clinical performance in Korean medical schools.

Yeungnam University has also conducted regional consortiums in training and its evaluation of clinical performances bi-annually. The consortium is primarily intended for educational feedback, so after conducting a similar evaluation of the KMLE, not only individual student feedback but also school feedback will be obtained.

The clinical skills test is consisting of the objective structured clinical examination (OSCE) and clinical performance examina-

tion (CPX). In evaluating communication skills and attitude by the standardized patient (SP), Yeungnam University employs five to six items that are the same as the physician-patient interaction (PPI) scales on the KMLE. These items include effective questioning, careful listening, a patient-centered attitude, using words precisely, and rapport building. In the physical examination station, another item has been included to assess a good manner during physical examination. Even though the six items of the PPI evaluation are based on global rating forms, the assessment guidelines and evaluation criteria for each question are quantitative rather than qualitative. They endeavor to establish an objective evaluation according to the checklist developed for this purpose. Faculty raters evaluate technical skills; SP raters evaluate PPI for the same students. This evaluation process is based on multiple studies conducted on the accuracy of information provided by trained SP raters and derived from SP assessments [1-3].

In general, the degree of agreement between faculty and SP scoring is 80% to 100%. However, research has shown inconsistencies in SP evaluations in Korea for 2,000, although SP score are higher than faculty scores. For example, in the study of Park et al. [4], the degree of agreement among the evaluations was 71% to 82%; the correlation among those evaluations appeared to have a 36% to 42% reliability, and the actual correlation itself was 0.60 to 0.65. The study of Kim et al. [5] regarding 14 evaluation items shows a significant difference between faculty and SP evaluations. In other studies, the correlation between faculty raters and SP raters for PPI was 0.54, which was lower than that in other regions [6,7]. Numerous researchers agree that the differences among the raters are due to a lack of explicit criteria and training experience, the rater's level of fatigue, and the number of items and rating scales [8,9].

Since 2010, however, there has not been enough exploratory research into the characteristics of these assessors or their reliability. This is the primary reason for the current study—to remedy evaluation methods, assessment skills, and training for raters. To accomplish reliable PPI evaluations, we compared faculty evaluations using the Yeungnam University Scale (YUS), analytic global rating scale (AGRS), and holistic rating scale (HRS) [10] with SP evaluation scores. In addition, research was conducted on the differences in faculty specialty and rater's experiences, as well as the rater's level of fatigue as time passes.

The research problems of this study are as follows: are there any differences in PPI scores between faculty raters and SP raters; what are the differences in PPI scores based on the faculty raters' specialty? What are the differences in PPI scores according to the raters' evaluation experience? Are the scores consistent from beginning to end during long evaluation times?

Materials and methods

1. Procedure and participants

The present study protocol was reviewed and approved an exemption of informed consent by the Institutional Review Board of Yeungnam University Hospital (IRB No: 7002016-E-2015-001).

For this study, we analyzed the PPI evaluation in the CPX that was conducted among third-year students at Yeungnam University College of Medicine in 2013. The evaluation was conducted on December 16 and 17. In total, 72 students were enrolled; there were 51 males (70.8%) and 21 females (29.2%) whose average age was 22.45 years. The group was subdivided randomly into 12 teams consisting of six members each. A total of 12 stations were set up to evaluate the CPX over 2 days; two copies of six stations were assigned each morning and afternoon.

To assess the reliability of the evaluation process, faculty raters and SP raters conducted the PPI evaluation simultaneously in each station; 48 faculty raters and 24 SP raters took part in the evaluations. Before the examination, all raters were trained for 2 hours on the general assessment process and the criteria for scoring assigned stations. We analyzed the evaluation results of eight stations, excluding four stations that did not have a complete scoring process. Consequently, we analyzed the final data based on evaluations conducted by 32 faculty and 16 SP raters. Of the 32 faculty raters, 12 evaluated stations related to their specialty while the other 20 evaluated stations out of their specialty. Nine had majored in basic medical science and 23 in clinical medicine. On average, faculty raters had participated in evaluation 5.13 times, while the average evaluation experience of SP raters was 7.44 times. Finally, the eight stations we examined included: jaundice, antenatal care, adult immunization, drinking problems, low back pain, micturition disorder, convulsion in childhood, and mastodynia/breast mass.

2. Research instruments

For this study, three rating scales were employed. The first scale, the YUS, was developed by the Daegu-Gyeungbuk regional consortium in Korea and revised by the Yeungnam University College of Medicine. The YUS consists of six items worth four points each; effective questioning, careful listening, patient-centered attitude, using words precisely, rapport building, and a good manner during physical examination. Also, the uniqueness of this rating scale is that it is similar to the global rating; however, it has been developed as a quantitative criterion to ensure evaluation objectivity. Table 1 shows an example of YUS. As a second scale, we employed four questions from AGRS developed by Hodges and

McIlroy [10]. The AGRS consists of the following four items worth five points each; response to patient's feelings and needs, degree of coherence in the interview, verbal expression, and non-verbal expression. This scale used qualitative criteria to assess the above items. Finally, the one-item HRS was employed as an additional question for the overall assessment of the knowledge and skills. Table 2 shows an example of AGRS and HRS.

3. Analysis

Regarding the 72 students' clinical performance, we compared differences in final scores using the rater's assessment expertise, rating experience, and the level of fatigue. Initially, all test items used T-score as a standard; the scores of YUS, AGRS, and HRS were compared with the average scores of the overall test items. In addition, we compared the faculty rater's scores with those of the SP raters for each of the evaluations. We also compared scores of the faculty raters who assessed stations related to their specialty with the scores of those who assessed stations out of their specialty and the differences between basic medical science and clinical medicine specialties. We proceeded with comparing score differences based on rater evaluation experience divided into three categories; less than three times, three to 10 times, and more than 10 times.

Furthermore, we examined the correlation coefficient based on the SP rater's level of fatigue. To test these group differences, the T-test and F-test were used. To verify differences based on SP rater fatigue, we compared the results of SPs who were assigned during the morning or afternoon with those who participated all day. Moreover, we looked at the correlation differences among all scales.

Table 1. An example of evaluation standards of Yeungnam University Scale

Item	Standard
1. He (or she) asked me something well and effectively	· Effective questions: open questions, confirmation questions, summary of the dialogues
	· Avoid questions: leading question, double meaning questions
	4. Excellent: asked all effective questions without any avoid questions
	3. Good: asked 2 effective questions and avoid questions
	2. Normal: asked 1 effective question or not asked any effective and avoid questions
	1. Not sufficient: not asked any effective question and asked avoid questions
2. He (or she) listened carefully	· Verbal response, listening attitude, eye contact, not take any speaking
	4. Excellent: used 4 actions very well
	3. Good: used 2 or 3 actions of all
	2. Normal: used 1 action or not used any actions and not used any opposite actions
	1. Not sufficient: used 4 opposite actions

Results

A comparison of raters' scores based on the evaluation tools for each item and the inter-rater correlation coefficient are shown in Table 3 and Fig. 1. The results show significant differences in scores between faculty and SP raters according to the evaluation tools. The scores of SP raters were significantly higher than those of the faculty raters; YUS ($t=-18.23$, $p<0.001$), AGRS ($t=-10.24$, $p<0.001$), and HRS ($t=-3.20$, $p<0.01$), respectively. Correlation coefficients between the faculty and SP raters were the following: YUS, 0.49 ($p<0.001$); AGRS, 0.34 ($p<0.01$); and HRS, 0.40 ($p<0.001$), respectively. These results indicate a significant correlation among the raters' scores based on each evaluation tool.

The results of the score differences according to the major congruence of the raters who assessed items related to their specialty and those who assessed items out of their specialty are in Table 4. There was a significant difference between the YUS ($t=-3.61$, $p<0.001$) and the AGRS ($t=-2.73$, $p<0.01$). The scores of the raters who assessed stations related to their specialty were lower than those of the raters who assessed stations out of their area of specialty. In the case of HRS, there was no significant difference.

Table 5 shows the results of score differences according to raters' evaluation experiences. In the YUS and AGRS, significant dif-

Table 2. An example of evaluation standards of analytic global rating scale and holistic rating scale

Example	Rating scale				
	1	2	3	4	5
Response to patient's feelings and needs (empathy)	Does not respond to obvious patient cues (verbal and non-verbal) and/or responds inappropriately		Responds to patient's needs and cues, but not always effectively		Responds consistently in a perceptive and genuine manner to the patient's needs and cues
Degree of coherence in the interview	No recognizable plan to the interaction; the plan does not demonstrate cohesion or the patient must determine the direction of the interview		Organizational approach is formulaic and minimally flexible and/or control of the interview is inconsistent		Superior organization, demonstrating command of cohesive devices flexibility, and consistent control of the interview

https://doi.org/10.12701/yujm.2020.00423

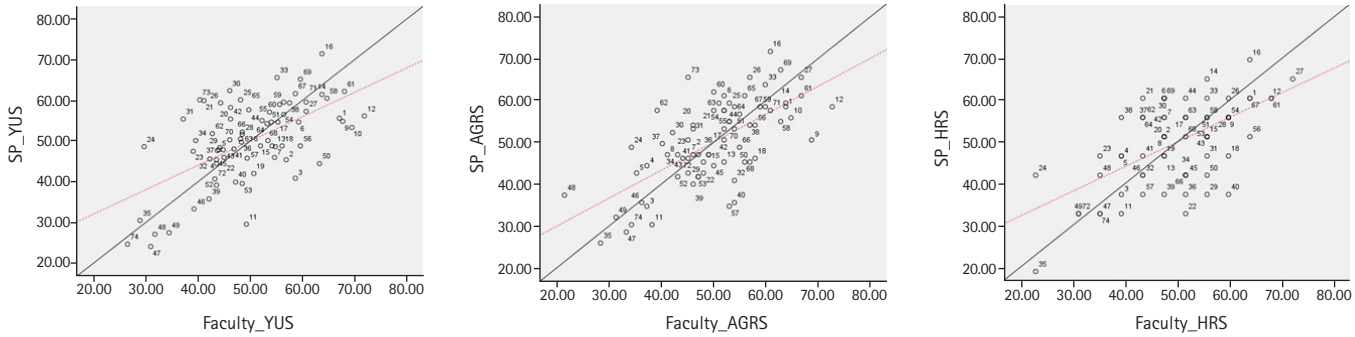

Fig. 1. X-Y scatter plot with scores of faculty and standardized patient. SP, standardized patient; YUS, Yeungnam University Scale; AGRS, analytic global rating scale; HRS, holistic rating scale.

Table 3. Comparisons of rating scales and Pearson's correlation coefficient of raters

Case	Rating scale	Rater						*t*	Pearson's correlation coefficient
		Faculty (n = 32)		SP (n = 16)		Total (n = 48)			
		n	Mean (SD)	n	Mean (SD)	n	Mean (SD)		
A	YUS	4	50.39 (8.41)	2	56.33 (2.53)	6	53.42 (6.82)	−5.70[c]	0.34[b]
	AGRS	4	49.81 (9.34)	2	55.49 (7.79)	6	52.71 (9.01)	−3.98[c]	−0.34[b]
	HRS	4	51.79 (10.50)	2	54.29 (9.28)	6	53.06 (9.94)	−1.51	−0.28[a]
B	YUS	4	43.98 (8.33)	2	53.20 (6.15)	6	48.63 (8.63)	−7.62[c]	0.26[a]
	AGRS	4	45.48 (5.96)	2	47.44 (12.35)	6	46.47 (9.73)	−1.23	0.12
	HRS	4	49.13 (9.45)	2	46.22 (12.90)	6	47.67 (11.37)	1.56	0.32[b]
C	YUS	4	45.34 (7.48)	2	50.24 (7.59)	6	47.85 (7.90)	−3.91[c]	0.36[b]
	AGRS	4	45.77 (7.36)	2	48.28 (10.30)	6	47.05 (9.04)	−1.70	0.27[a]
	HRS	4	45.89 (9.30)	2	46.66 (11.00)	6	46.28 (10.17)	−0.45	0.39[c]
D	YUS	4	46.32 (9.73)	2	53.61 (4.48)	6	50.50 (8.03)	−5.16[c]	−0.11
	AGRS	4	46.71 (10.35)	2	51.54 (6.47)	6	49.48 (8.65)	−3.05[b]	−0.09
	HRS	4	47.79 (10.50)	2	47.75 (7.21)	6	47.76 (8.73)	0.02	0.06
E	YUS	4	47.61 (6.58)	2	55.35 (3.96)	6	51.64 (6.61)	−8.39[c]	0.26[a]
	AGRS	4	49.48 (6.53)	2	55.19 (6.27)	6	52.46 (6.99)	−5.31[c]	0.21
	HRS	4	51.21 (9.84)	2	55.37 (7.09)	6	53.38 (8.74)	−2.87[b]	0.32[b]
F	YUS	4	45.19 (5.67)	2	54.93 (3.12)	6	50.09 (6.68)	−12.88[c]	0.35[b]
	AGRS	4	47.59 (7.27)	2	56.85 (4.78)	6	52.25 (7.68)	−9.11[c]	0.03
	HRS	4	48.03 (9.10)	2	58.43 (4.07)	6	53.26 (8.74)	−8.91[c]	0.14
G	YUS	4	41.50 (8.64)	2	49.77 (4.44)	6	45.09 (8.21)	−7.05[c]	0.23
	AGRS	4	43.21 (8.62)	2	49.73 (5.35)	6	46.04 (8.04)	−5.27[c]	0.24
	HRS	4	45.38 (10.73)	2	46.84 (7.35)	6	46.01 (9.41)	−0.92	0.45[c]
H	YUS	4	49.22 (6.31)	2	54.24 (3.56)	6	51.87 (5.62)	−5.69[c]	0.17
	AGRS	4	51.91 (5.19)	2	54.61 (3.98)	6	53.34 (4.77)	−3.42[c]	0.18
	HRS	4	53.41 (8.84)	2	51.02 (8.10)	6	52.14 (8.51)	1.67	0.06
Total	YUS	4	46.13 (8.12)	2	53.57 (5.18)	6	49.93 (7.73)	−18.23[c]	0.29[c]
	AGRS	4	47.44 (8.08)	2	52.48 (8.38)	6	50.01 (8.61)	−10.24[c]	0.14[b]
	HRS	4	49.05 (10.09)	2	50.95 (9.78)	6	50.02 (9.98)	−3.20[b]	0.20[c]

SP, standardized patients; SD, standard deviation; YUS, Yeungnam University Scale; AGRS, analytic global rating scale; HRS, holistic rating scale; A, jaundice; B, antenatal; C, adult immunization; D, drinking problem; E, low back pain; F, micturition disorder; G, convulsion in childhood; H, mastodynia, breast mass.
[a] $p<0.05$. [b] $p<0.01$. [c] $p<0.001$.

ferences based on the faculty's evaluation experience were observed. It was observed that the scores from those who had participated 3–10 times group were lower than those from the other two groups. In all scales, there were significant differences among the SP raters. However, for the SP raters, scores from those who had participated 3–10 times were higher than those from the other groups.

The reviewed correlation between three rating scales in each group of raters is shown in Table 6. In the results of the correlation between the rating scales, the following were observed; 0.66 to 0.83 in faculty ($p < 0.001$) and 0.68 to 0.82 in SP ($p < 0.001$).

To find the impact of the raters' level of fatigue on evaluation outcome, we compared the scores given by SP raters in the morning and afternoon. The results are shown in Table 7. The correlation between the YUS and AGRS and between the YUS and HRS were significantly lower in the afternoon than in the morning.

Discussion

Analyzing the differences between the evaluation results of the three assessment tools and the raters' characteristics, we have explored factors that may affect PPI scores. This study was conducted to minimize the effects of co-factors to secure fairness. Therefore, this study tried to find out the difference between evaluation results according to the characteristics of evaluators and evaluation tools and discuss the implications of each result.

The result of this study showed that there was a slight difference in PPI scores between faculty raters and SP raters. In general, when there was a significant difference, the evaluation scores of SPs were significantly higher than that of the faculty. These results were similar to previous studies due to the SP role, which was not primarily for evaluation purposes but rather for constructive educational feedback. Hence, SPs are not comfortable with giving low scores. However, faculty raters deal with subjects with which they are familiar; consequently, they can be better focused and stricter in their evaluation [11-13]. In addition, concerning the YUS, AGRS, and HRS scores, Pearson's correlation coefficient between the faculty and SP raters was 0.34 to 0.49. In the study of Kwon et al. [6], the results of the CPX concerning technical skill items-history taking, physical examination, and patient education-showed correlation coefficient from 0.69 to 0.91. In contrast, the correlation for the patient-doctor relationship was somewhat low; 0.09 to 0.51. The reasons for these results could be found in ambiguous applied standards, the level of raters' fatigue, and lack of training [6-9,14]. For additional improvement of the evaluation process, this study recommends periodic evaluation of raters and further training tasks.

The SPs involved in this study had participated in PPI evaluations using the YUS an average of 7.44 times and had at least 3 to 12 hours of training according to the training guidelines set by the consortium. However, in the training for the AGRS and HRS, the instruction portion involved less than three hours of training. Furthermore, the raters' fatigue from having to assess on all three scales could affect the evaluation outcome. In the case of faculty raters, even though they had an average of 5.13 evaluation experi-

Table 4. Comparisons of scores according to the congruence of the raters' specialty

Rating scale	Specialty		Total (n = 32)	*t*
	Congruence (n = 12)	Incongruence (n = 20)		
YUS	44.36 (5.94)	46.74 (8.68)	46.13 (8.12)	−3.61[a]
AGRS	46.00 (6.91)	47.95 (8.39)	47.44 (8.08)	−2.73[b]
HRS	48.61 (10.26)	49.20 (10.05)	49.05 (10.09)	−0.60

Values are presented as mean (standard deviation).
YUS, Yeungnam University Scale; AGRS, analytic global rating scale; HRS, holistic rating scale.
[a] $p < 0.001$. [b] $p < 0.01$.

Table 5. Comparisons of scores between evaluation experiences of faculty raters and standardized patients

Rating scale	Rater	Evaluation experience (time)						Total		F	Scheffé
		< 3		3–10		> 10					
		n	Mean (SD)	n	Mean (SD)	n	Mean (SD)	n	Mean (SD)		
YUS	Faculty	6	48.49 (7.48)	20	44.35 (7.64)	6	49.03 (8.83)	32	46.13 (8.12)	19.78[a]	3,1 > 2
	SP	3	48.57 (5.09)	7	55.57 (3.71)	6	51.89 (5.84)	16	53.57 (5.18)	39.68[a]	2 > 1,3
	Total	9	49.68 (6.81)	27	49.37 (8.33)	12	50.94 (7.09)	48	49.93 (7.73)	4.39[b]	3 > 2
AGRS	Faculty	6	51.66 (7.83)	20	46.04 (7.26)	6	50.49 (9.55)	32	47.44 (8.08)	14.11[a]	3,1 > 2
	SP	3	51.22 (6.00)	7	55.29 (6.69)	6	49.78 (9.87)	16	52.48 (8.38)	30.19[a]	2 > 1,3
	Total	9	47.29 (7.20)	29	50.18 (8.38)	10	50.01 (9.76)	48	50.01 (8.61)	0.4	
HRS	Faculty	6	49.96 (10.30)	20	48.18 (10.08)	6	50.87 (9.69)	32	49.05 (10.09)	3.25[b]	
	SP	3	49.80 (7.18)	7	53.48 (8.90)	6	49.64 (10.88)	16	50.95 (9.78)	19.25[a]	2 > 1,3
	Total	9	48.54 (9.15)	27	50.55 (9.92)	12	50.04 (10.50)	48	50.02 (9.98)	3.17[b]	2 > 1

SD, standard deviation; YUS, Yeungnam University Scale; SP, standardized patient; AGRS, analytic global rating scale; HRS, holistic rating scale.
[a] $p < 0.001$. [b] $p < 0.05$.

https://doi.org/10.12701/yujm.2020.00423

Table 6. Comparisons of correlation between three rating scales in each group of raters

Pearson's correlation coefficient	Faculty (n = 32)	Standardized patient (n = 16)	Total (n = 48)
YUS-AGRS correlation	0.83[a]	0.77[a]	0.80[a]
AGRS-HRS correlation	0.75[a]	0.82[a]	0.77[a]
YUS-HRS correlation	0.66[a]	0.68[a]	0.62[a]

YUS, Yeungnam University Scale; AGRS, analytic global rating scale; HRS, holistic rating scale.
[a] $p < 0.001$.

Table 7. Comparisons of correlations between morning and afternoon in standardized patient raters

Pearson's correlation coefficient	Morning (n = 16)	Afternoon (n = 16)	Total (n = 32)	*t*
YUS-AGRS correlation	0.82 (0.08)	0.71 (0.10)	0.76 (0.11)	3.32[a]
AGRS-HRS correlation	0.80 (0.12)	0.76 (0.13)	0.78 (0.12)	0.75
YUS-HRS correlation	0.72 (0.14)	0.59 (0.10)	0.65 (0.14)	2.74[b]

Values are presented as mean (standard deviation).
YUS, Yeungnam University Scale; AGRS, analytic global rating scale; HRS, holistic rating scale.
[a] $p < 0.01$. [b] $p < 0.05$.

ences, their previous experience was limited only to evaluating technical skills; thus, they did not have proper training in conducting PPI evaluations. Since the pre-evaluation training for this study was completed in only one hour, overall training for the evaluation process was insufficient. There is another possibility that the raters' fatigue level was high because we conducted the technical skill and PPI evaluations together. The fact that the reliability among the scales evaluated in the afternoon is lower than that in the morning can be interpreted as an increase in the fatigue of the evaluator and a decrease in concentration. Therefore, it is necessary to arrange evaluators separately in the afternoon or provide sufficient rest time.

There were significant differences in the scores related to the raters' specialty and in scores between basic medical science and clinical medicine. Regarding the evaluation experience, the scores showed significant differences among each faculty and SP group. In the case of the faculty, the scores of evaluators with 3–10 evaluation experiences were significantly lower than those with less than three or more than 10 evaluation experiences. However, in the case of SPs, the evaluation scores of evaluators with 3–10 experiences were significantly higher than those with less than three or more than 10 evaluation experiences. These results are the same for all YUS, AGRS, and HRS evaluations.

According to these results, it was found that the faculty members were more thoroughly assessed if their majors matched, or if they had more than three or less than 10 evaluation experiences. However, SPs appeared to be less thorough, as they became accustomed to evaluation and became more involved when they had three to 10 evaluation experiences. To offset these differences and increase reliability, we need to strengthen the rater's competences through systematic strategies during their training and feedback. In addition, it is necessary to systematically apply a statistical method to offset the score difference between evaluators. Accordingly, in medical schools, it is necessary to train or acquire evaluation experts or analysis experts, and it is necessary to secure an evaluation system or an evaluation support team that is capable of systematic evaluation support.

Besides, in Korea, technical skills are assessed by faculty, but humanistic skills are assessed by SPs. The faculty raters tend to over-emphasize the importance of technical skills and, thus, evaluate them more strictly. However, they are inclined to be laxer and more generous when evaluating humanistic skills because they would feel awkward doing otherwise [15]. Moreover, Park et al. [7] reported lower PPI scores from SP evaluators through comparative studies on the accuracy of scoring by faculty and SPs on the CPX, confirming that there are also differences between evaluators who are simply observant or participatory. This means that it may also depend on the extent of the evaluator's subjective intervention, and therefore, accurate feedback on humanistic skills can be hindered in the process of evaluation feedback and retraining. Hence, faculty raters need to recognize the importance of humanistic skills and feedback for their results. Separate sessions of technical and humanistic skills on the CPX needs to be examined closely; both adequacy and reliability should be reviewed.

This study also tried to explore the appropriateness of the evaluation tool. Numerous previous studies on the subject report that the global rating is similar to the checklist or even superior in validity and reliability. Furthermore, to evaluate students' performance, they recommend using the checklist and the global rating scale together [11,16-18]. The checklist is primarily known for the evaluation of technical skills, including procedural performance. However, the global rating scale is known as an efficient tool to measure a student's attitudes and PPI, communication skills, and possible expertise [16,18-23]. Therefore, as far as humanistic skills are concerned, it is appropriate to employ the global rating scale for the PPI.

The worksheet has been developed as a checklist, although the PPI evaluation scales include global rating scale items at the Yeungnam University College of Medicine. This is because the checklist is superior in apparent objectivity and ease of use [24]. However, the checklist form of the worksheet may be more appropriate for the technical skills station, and the global rating form may be more appropriate for attitudes in the PPI station [25]. Generally, raters who use the checklist are inclined to give the same scores to

the students whether they performed well [26]. In this study, one of the SPs reported that there were students who were not satisfied with the overall physician-patient relationship, although she gave high scores for a lack of demerit factors according to the checklist.

In addition, students tend to memorize the checklist rather than practice technical skills in the process of preparing for the OSCE [27]. In this context, the phenomenon diminishes the value of education; specifically, it reduces the validity of assessing students' competence [12]. It is therefore imperative that we provide improved assessment tools for students so that the skills are embodied rather than a simple recitation of a memorized PPI checklist. Furthermore, it is necessary that students internalize the correct attitude in their practice. Thus, we propose a further study on how to improve evaluation tools and standards as a form of the AGRS.

In this study, SP assesses an average of six students at one station. We compared the correlations among the individual assessment scales, dividing the process into two different time periods-morning and afternoon-to find out the potential impact that raters' fatigue can have on the scores. We found a significant difference in the correlation between YUS and AGRS, and between YUS and HRS. Scores given in the afternoon session were lesser reliable than those given in the morning, which is statistically significant. Indeed, it is hard to stay focused on evaluations for a long time. The rater's fatigue, caused by the multitude of evaluation items and the number of students to evaluate, is responsible for the failure to maintain consistency [28-30]. To enhance the degree of consistency among raters, a strategy to lengthen the training period but reduce the time that raters spend on assessments could be proposed. Furthermore, specific ways could be proposed to use professional SPs more efficiently. As described above, SP evaluators received 3 to 12 hours of pre-training for CPX evaluation whereas faculty members received only one hour of pre-training. The reason for the low correlation between evaluation scores between faculty members and SP evaluators may also be because the standardized evaluator education was not conducted equally. It is suggested that systematic training is required for all evaluators to increase the reliability of evaluation scores.

This study based on one CPX assessment conducted at a school, and it is focused on the evaluation of the patient-doctor relationship rather than the clinical performance process. Due to temporal and financial limitations in the actual evaluation, two SPs participated in one clinical item evaluation, while four professors participated in the same evaluation. Therefore, the decrease in reliability for the evaluation tools of SP evaluators in the afternoon may be due to the increased fatigue.

Although the correlation was a little low, it was considered that the correlation between faculty and SP evaluators was higher when YUS was used than when AGRS or HRS were used. In the actual evaluation situation, ease of evaluation is considered as an important factor, and the evaluation tools with high validity and reliability with a few evaluation items are preferred. It is recommended to use HRS for the ease of evaluation, but it can be seen that the use of AGRS or YUS that comprises four to six items may be more appropriate to compensate for the limitations regarding single-item measurement. The most appropriate evaluation tool should be suitable in the form and number of evaluation items, and both evaluation validity and reliability must be satisfied. Therefore, it is necessary to continuously study which evaluation tools are useful for evaluating according to various evaluation situations and clinical presentations.

The limitations of this study are as follows: First, this cannot be generalized because it based on the experience of a single school. Second, it is a retrospective study, not a study designed to identify only the factors that affect PPI scores. Third, this is a separate analysis of the PPI scores only, which are parts of the station for CPX. Therefore, it did not reflect the impacts of the characteristics (e.g., difficulty level, etc.) of each station. Finally, it did not reflect the personal characteristics (e.g., beliefs, emotional state, fatigue level, etc.) of raters on the day of assessment.

In conclusion, the reliability of PPI scores on the CPX was found to be significantly affected by evaluator factors as well as the type of scale used. There is a need for a further study to establish guidelines for evaluating PPI on the CPX and to offer appropriate assessment tools, an ideal number of items, raters' qualification, and ideal length of evaluation time.

Acknowledgments

Conflicts of interest

No potential conflict of interest relevant to this article was reported.

Funding

This research was supported by a grant of Yeungnam University Medical Center (2011).

Author contributions

Conceptualization and Validation: all authors; Project administration: KSL, YHL; Funding acquisition: YHL; Data curation and Formal analysis: KHC, YSP; Writing-original draft: YSP, KSL; Writing-review & editing: all authors.

https://doi.org/10.12701/yujm.2020.00423

ORCID
Young Soon Park, https://orcid.org/0000-0002-3644-8793
Kyung Hee Chun, https://orcid.org/0000-0002-5351-0376
Kyeong Soo Lee, https://orcid.org/0000-0001-8183-9462
Young Hwan Lee, https://orcid.org/0000-0001-8377-5802

References

1. Elliot DL, Hickam DH. Evaluation of physical examination skills: reliability of faculty observers and patient instructors. JAMA 1987;258:3405–8.
2. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. The accuracy of standardized patient presentation. Med Educ 1991;25:100–9.
3. Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, Barrows HS. Standardized (simulated) patients' accuracy in recording clinical performance check-list items. Med Educ 1992;26:99–104.
4. Park H, Lee J, Hwang H, Lee J, Choi Y, Kim H, et al. The agreement of checklist recordings between faculties and standardized patients in an objective structured clinical examination (OSCE). Korean J Med Educ 2003;15:143–52.
5. Kim JJ, Lee KJ, Choi KY, Lee DW. Analysis of the evaluation for clinical performance examination using standardized patients in one medical school. Korean J Med Educ 2004;16:51–61.
6. Kwon I, Kim N, Lee SN, Eo E, Park H, Lee DH, et al. Comparison of the evaluation results of faculty with those of standardized patients in a clinical performance examination experience. Korean J Med Educ 2005;17:173–84.
7. Park J, Ko J, Kim S, Yoo H. Faculty observer and standardized patient accuracy in recording examinees' behaviors using checklists in the clinical performance examination. Korean J Med Educ 2009;21:287–97.
8. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63–7.
9. De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. Acad Med 1997;72(10 Suppl 1):S85–7.
10. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37:1012–6.
11. Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H. Assessing students' communication skills: validation of a global rating. Adv Health Sci Educ Theory Pract 2008;13:583–92.
12. Heine N, Garman K, Wallace P, Bartos R, Richards A. An analysis of standardised patient checklist errors and their effect on student scores. Med Educ 2003;37:99–104.
13. McLaughlin K, Gregor L, Jones A, Coderre S. Can standardized patients replace physicians as OSCE examiners? BMC Med Educ 2006;6:12.
14. Tamblyn RM, Klass DJ, Schnabl GK, Kopelow ML. Sources of unreliability and bias in standardized-patient rating. Teach Learn Med 1991;3:74–85.
15. Domingues RC, Amaral E, Zeferino AM. Global overall rating for assessing clinical competence: what does it really show? Med Educ 2009;43:883–6.
16. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. Adv Health Sci Educ Theory Pract 1996;1:227–33.
17. Nielsen DG, Gotzsche O, Eika B. Objective structured assessment of technical competence in transthoracic echocardiography: a validity study in a standardized setting. BMC Med Educ 2013;13:47.
18. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998;73:993–7.
19. Cohen R, Rothman AI, Poldre P, Ross J. Validity and generalizability of global ratings in an objective structured clinical examination. Acad Med 1991;66:545–8.
20. LeBlanc VR, Tabak D, Kneebone R, Nestel D, MacRae H, Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. Am J Surg 2009;197:96–101.
21. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med 1999;74:1129–34.
22. Turner K, Bell M, Bays L, Lau C, Lai C, Kendzerska T, et al. Correlation between global rating scale and specific checklist scores for professional behaviour of physical therapy students in practical examinations. Educ Res Int 2014;2014:219512.
23. Wilkinson TJ, Fontaine S. Patients' global ratings of student competence: unreliable contamination or gold standard? Med Educ 2002;36:1117–21.
24. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ 2015;49:161–73.
25. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004;38:199–203.
26. Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M. Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. Simul Healthc 2013;8:148–54.
27. van Luijk SJ, van der Vleuten CPM, van Schelven SM. Observer

and student opinions about performance-based tests. In: Bender W, Hiemstra RJ, Scherpbier AJ, Zwierstra RP, editors. Teaching and assessing clinical competence. Groningen: Boekwerk Publications; 1990. p. 497–502.
28. Klein SP, Stecher BM, Shavelson RJ, McCaffrey D, Ormseth T, Bell RM, et al. Analytic versus holistic scoring of science performance tasks. Appl Meas Educ 1998;11:121–37.
29. Regehr G, Freeman R, Hodges B, Russell L. Assessing the generalizability of OSCE measures across content domains. Acad Med 1999;74:1320–2.
30. Choi JY, Jang KS, Choi SH, Hong MS. Validity and reliability of a clinical performance examination using standardized patients. J Korean Acad Nurs 2008;38:83–91.

https://doi.org/10.12701/yujm.2020.00423