Original Article

Inter-Rater Reliability of Grade Evaluation of Post-Clinical Clerkship (Post-CC) OSCE Based on Kappa Coefficient and Agreement Rates

Mitsuru Takayama[1, 2]* Akiko Nakada[1, 3] and Naoki Hiroi[1, 3]

[1] Department of Medical Education, Toho University Graduate School of Medicine, Tokyo, Japan
[2] Department of Pediatric Nursing, Faculty of Nursing, Toho University, Tokyo, Japan
[3] Center for Medical Education, Faculty of Medicine, Toho University, Tokyo, Japan

ABSTRACT

Introduction: Post-Clinical Clerkship (Post-CC) Objective Structured Clinical Examination (OSCE) has been implemented as a unified examination in all medical schools since fiscal 2020. In this study, differences in the grade evaluation of Post-CC OSCE made by faculty members in fiscal 2020 were investigated.

Methods: Grade evaluations of Post-CC OSCE, which was taken by 119 students and 36 faculty members, were analyzed. For the calculation of differences in the evaluation scores between two evaluator faculty members within a station, weighted kappa coefficient, Spearman's rank correlation coefficient, and simple percent agreement were used. The overall evaluation directly linked to a student's pass/fail was compared in each series using the Kruskal-Wallis test and $\chi 2$ test.

Results: The concordance rate of grade evaluation between 2 evaluator faculty members was low and a pass/fail judgment was disagreed on at 35% probability in "physical examination." There were no significant differences among the overall evaluations of each series. However, when the evaluations were categorized into 2 groups on the basis of the pass line (≥ 4 and < 4), there was a significant difference in the pass rate of each series and it ranged from 36.8% to 90.0%.

Conclusions: Disagreement of grade evaluation directly linked to pass/fail was noted between evaluator faculty members as well as between series. In addition to a review of the implementation and assessment methods of examination, it is important to construct a system capable of reviewing the evaluation after an examination.

Toho J Med 8 (2): 61–70, 2022

KEYWORDS: Post-Clinical Clerkship OSCE, interrater reliability, medical education, evaluator

Introduction

Objective Structured Clinical Examination (OSCE) is an attitude/skill assessment method developed by Harden et al[1] that has become widely used as an innovative method for objectively evaluating clinical competence. In Japan, it has gradually spread since Kawasaki Medical School introduced it as a basic clinical competence assessment method

*Corresponding Author: Mitsuru Takayama, 5-21-16 Omorinishi, Ota-ku, Tokyo, 143-0015, Japan, tel: 03–3762–9881
e-mail: mitsuru.takayama@med.toho-u.ac.jp

DOI: 10.14994/tohojmed.2021-015

Received Oct. 19, 2021: Accepted Dec. 16, 2021
Toho Journal of Medicine 8 (2), June 1, 2022.
ISSN 2189–1990, CODEN: TJMOA2

in 1993,[2] and it was implemented in 2005 as a common test to evaluate students before clinical training. The Physicians Subcommittee of the Ministry of Health, Labour and Welfare Medical Ethics Council proposed that OSCE be considered an official examination in the future.[3] Furthermore, a trial of Post-Clinical Clerkship (Post-CC) OSCE was initiated by the Common Achievement Tests Organization in 2016 aimed at securing the clinical competence of medical students at the time of graduation at a certain level or higher throughout the country, which was officially implemented in a way unified by the Common Achievement Tests Organization in 2020.[4] Post-CC OSCE is a high-stakes test performed to comprehensively evaluate the proficiency of "knowledge and skills which a medical practitioner should possess" as stipulated in Article 9 of the Medical Practitioners' Act.

At present, no practical test is included in the National Examination for Medical Practitioners, so it is important to secure a high reliability of this examination, but reliability may be lost because of the quality of raters. All medical schools in Japan have implemented Post-CC OSCE in fiscal 2020,[4] but only a few studies reported the reliability of the evaluation of Post-CC OSCE. Makuuchi et al.[5] investigated concurrences and differences in the evaluation between faculty members and standardized patients (SP), but they did not mention the variation of the evaluation among faculty members. Hara et al.[6] investigated the reliability of Post-CC OSCE using the generalizability theory and stated that increasing raters, tasks, and assessment items are necessary to secure sufficient reliability. Of course, the objectivity of the evaluation is necessary because Post-CC OSCE is an examination, whereas minimizing differences in the evaluation is necessary when multiple raters perform an assessment. In a discussion about the reliability of medical and dental OSCE, it has been pointed out that the score of each assessment item is frequently influenced by the subjectivity of raters, resulting in the occurrence of variation in the evaluation among raters.[7–9]

Therefore, the objective of this study was to clarify differences in grade evaluation made by evaluator faculty members and investigate measures improving the assessment method.

Methods

Subjects and survey period

The Post-CC OSCE evaluation scores of 119 6th-year medical students of a university assessed by 36 raters in fiscal 2020 were analyzed. This examination comprised three tasks in total; six series were assigned to each task, and two raters were arranged per station (a room to assess clinical skills based on OSCE) (Fig. 1). One student was assessed by two raters simultaneously.

Post-CC OSCE assessment employed a rubric, and the following six items were assessed: A, consideration for/communication with patients; B, medical interview; C, physical examination based on diagnostic hypothesis; D, case presentation; E, clinical reasoning; and overall evaluation. The evaluation scale of 1-6 was used as the scores for analysis.

Statistical analysis

The grade evaluation using a rubric was regarded as a dependent variable, and grade evaluation-associated factors (series, time zone in which examination was performed, position of the rater) were regarded as independent variables. It is considered that the influence of the rate of coincidences cannot be eliminated by simply observing the percent agreement, and to increase the reliability of evaluation, it is desirable to use a "concordance rate (reproducibility)" in which coincidences were eliminated from the percent agreement.[10] Weighted kappa (κ) coefficient is an index that assesses the reproducibility between 2 raters in which the value is 0 when the actual percent agreement is equal to incidents, i.e., due to coincidences, and 1 when the agreement is complete. The guidelines concerning the level of κ coefficient at which interrater reliability is maintained are available. In this study, $\kappa \leq 0.2$, $0.2 < \kappa \leq 0.4$, $0.4 < \kappa \leq 0.6$, $0.6 < \kappa \leq 0.8$, and $0.8 < \kappa \leq 1.0$ were judged as Poor, Fair, Moderate, Good, and Very good, respectively.[10]

Confidence intervals for κ coefficient and correlation coefficient between the 2 raters may be relatively small when around 50% of individuals selected a particular option ("score rate"); however, confidence intervals may increase as the score rate approaches 0% or 100%. As such, it is recommended to indicate both the interrater percent agreement and either the κ value or correlation coefficient.[11] Following this, κ coefficient, Spearman's rank correlation coefficient, and interrater percent agreement were used in this study, when the evaluation score was compared by assessment item between the two raters within a station. Interrater percent agreement was calculated for the 6-grade evaluation as well as the 2-grade evaluation set based on the pass line of 4 (≥4 and <4) 4. Furthermore, Kruskal-Wallis test and $\chi 2$ test were used to

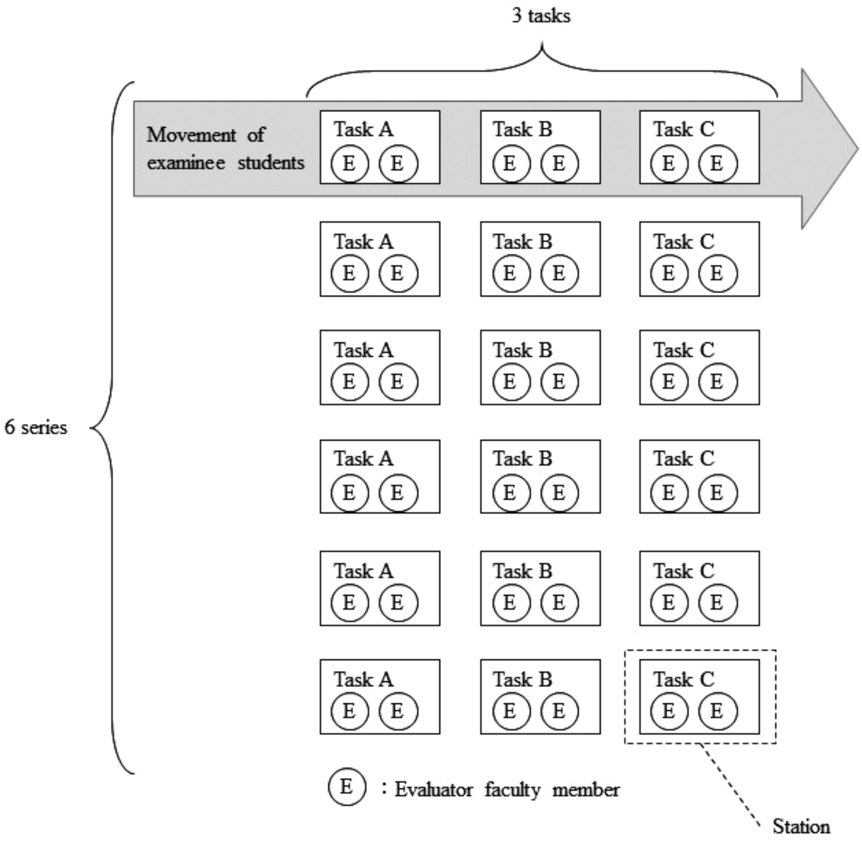


Fig. 1 Flow of examination

compare the mean overall evaluation score of three tasks by series. Since the overall evaluation score is used to determine whether the student has passed or failed, we selected the mean of the overall evaluation score for the three tasks as representative values. Kruskal-Wallis test was used to compare the differences in the scores for each assessment item based on the position of the raters. Bonferroni correction was performed as a post-hoc test. Mann-Whitney U-test was used to compare the differences in the scores for each assessment item based on the time of assessment. Statistical significance was set at $p < 0.05$. Since students with experience of repeating a grade were concentrated in a specific time zone, students with experience of repeating a grade were excluded from the examination-implemented time zone-based analysis.

Ethical considerations

This study was approved by the Ethics Committee of the Toho University School of Medicine (approval No. A20051_A19089).

To obtain consent from the students, an explanatory document was sent to the subjects by mail, and when a subject did not give consent, it was informed to the investigator. The same method was used to obtain consent from faculty members, who were raters, in which the mail to students was sent by an investigator who was not a faculty member of the medical school to minimize the pressure on students. Since the data contained students' records, the survey was initiated after a pass/fail judgment was made.

Results

Summary

Since all students and raters gave consent to participation in this survey, all evaluation scores were included in the analysis. In fiscal 2020, 119 students of a university medical school took Post-CC OSCE comprising 3 tasks, so that 357 examinations were performed in total and 2 raters simultaneously assessed these, so that 714 assessments were performed in total. The positions were professor in 12 raters (33.3%), associate professor in 13 (36.1%), and lecturer in 11 (30.6%), and 32 were male (88.9%) and 4 were female (11.1%). Ninety-four (79.0%) and twenty-five (21.0%) examinee students had no experience and had an experience of repeating a grade, respectively. The mean

Table 1 Summary of grade evaluation

	A Consideration/communication	B Medical interview	C Physical examination	D Case presentation	E Clinical reasoning	Overall evaluation
Mean	4.58	4.44	3.80	4.30	4.10	4.04
SD	0.34	0.35	0.42	0.39	0.46	0.44
Maximum value	5.50	5.50	4.83	5.17	5.33	5.17
75th PCTL	4.83	4.67	4.17	4.50	4.33	4.33
Median	4.50	4.50	3.83	4.33	4.17	4.00
25th PCTL	4.33	4.17	3.50	4.00	3.83	3.67
Minimum value	3.83	3.50	2.83	3.33	2.83	2.83

SD: Standard Deviation; PCTL: Percentile

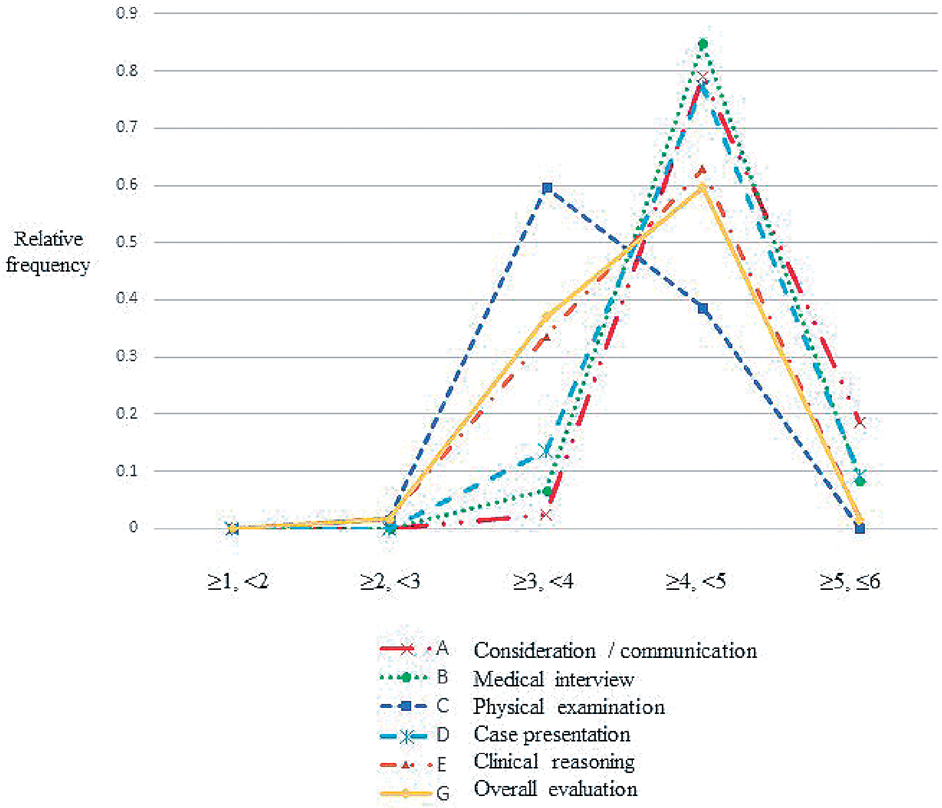


Fig. 2 Frequency distribution of mean score of all tasks by assessment item (score of each student is the mean of multiple tasks)
A: Consideration/communication
B: Medical interview
C: Physical examination
D: Case presentation
E: Clinical reasoning
G: Overall evaluation

score of each assessment item was 4.58 ± 0.34 in Item A, 4.44 ± 0.35 in Item B, 3.80 ± 0.42 in Item C, 4.30 ± 0.39 in Item D, and 4.10 ± 0.46 in Item E, and the overall evaluation was 4.04 ± 0.44 (Table 1, Fig. 2).

Intrastation difference between the two raters

The intrastation difference between the 2 raters was evaluated by assessment items of the examination performed 357 times (Table 2). κ coefficient of reproducibility

Table 2 Reproducibility between two raters within a station

		A	B	C	D	E	
		Consideration/ communication	Medical interview	Physical examination	Case presentation	Clinical reasoning	Overall evaluation
Weighted kappa coefficient †		0.18	0.22	0.22	0.22	0.27	0.27
Spearman's rank correlation coefficient		.256 *	.315 *	.315 *	.339 *	.424 *	.371 *
Percent agreement	6 grades	46.2%	47.1%	41.7%	48.5%	45.7%	49.9%
	Passing line	95.5%	86.8%	65.3%	85.4%	78.2%	82.4%

† κ ≤ 0.2: Poor, 0.2< κ ≤ 0.4: Fair, 0.4< κ ≤ 0.6: Moderate, 0.6< κ ≤ 0.8: Good, 0.8< κ ≤ 1.0: Very good
* p<0.001

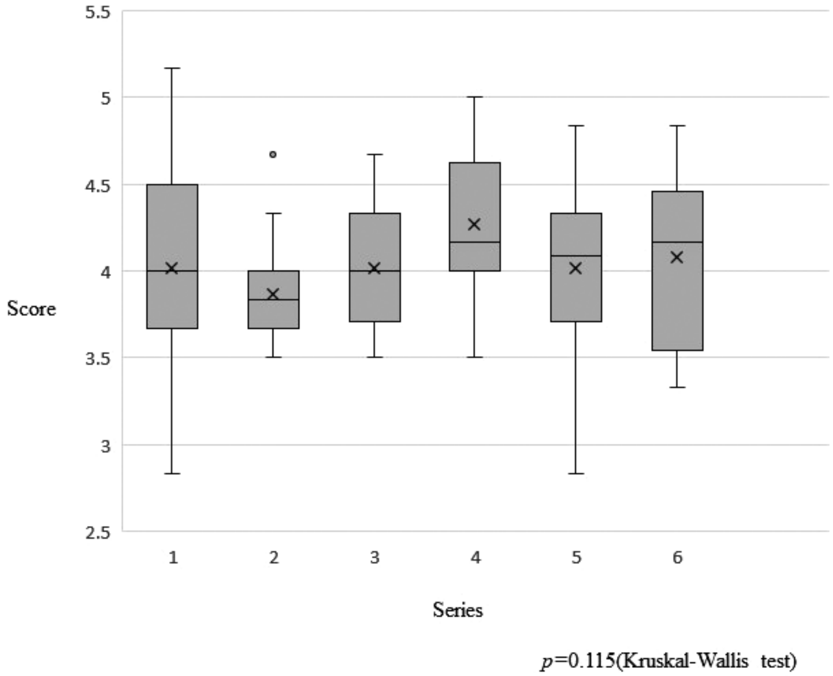


Fig. 3 Overall evaluation score by series

was the lowest in "Item A: consideration/communication" (κ = 0.18, Poor) and highest in "Item E: clinical reasoning" and "overall evaluation" (κ = 0.27, Fair). Spearman's rank correlation coefficient was the lowest in "Item A: consideration/communication" (r = 0.256, p < 0.001) and highest in "Item E: clinical reasoning" (r = 0.424, p < 0.001). The lowest percent agreement of the 6-grade Likert assessment was 41.7% (121/357) noted in "Item C: physical examination," and the highest was 49.9% (128/357) noted in "overall evaluation." However, when the 6-grade Likert was divided into two groups by the pass line, 4 or higher and below 4, and the percent agreement was calculated, the highest percent agreement, 95.5% (341/357), was noted in "Item A: consideration/communication," whereas the lowest percent agreement was noted in "Item C: physical examination," and it was 65.3% (233/357), showing that a pass/fail judgment was disagreed on between the two raters in many stations.

Series-associated difference in grade evaluation

The mean overall evaluation score of the three tasks that the 119 students took was calculated and is presented as a graph in Fig. 3. The score was the highest in Series 4, and it was 4.27 ± 0.41, whereas the score was the lowest in Series 2, and it was 3.87 ± 0.33, showing no significant difference between series (p = 0.115). However, when the overall evaluation score was divided based on the pass line, 4 or higher and below 4 (Table 3), the highest pass rate, 90.0% (18/20), was noted in Series 4, and the lowest pass rate, 36.8% (7/19), was noted in Series 2, showing a large difference, and the pass rate was significantly different between the series (p = 0.023).

Examination-implemented time zone-associated difference

The grade evaluation of 94 students with no experience of repeating a grade was divided based on the examination-implemented time zone into two groups: the former and the latter half time zones and the mean were calculated (Fig. 4). No significant difference was noted in any of the six assessment items between the former and the latter half groups (p = 0.207-0.955).

Table 3 Number of successful examinees (who passed) whose mean overall evaluation of all tasks they took was 4 or higher

Series	Successful examinees/ examinees	(Pass rate)
1	11/20	(55.0%)
2	7/19	(36.8%)
3	11/20	(55.0%)
4	18/20	(90.0%)
5	12/20	(60.0%)
6	14/20	(70.0%)
Total	73/119	(61.3%)

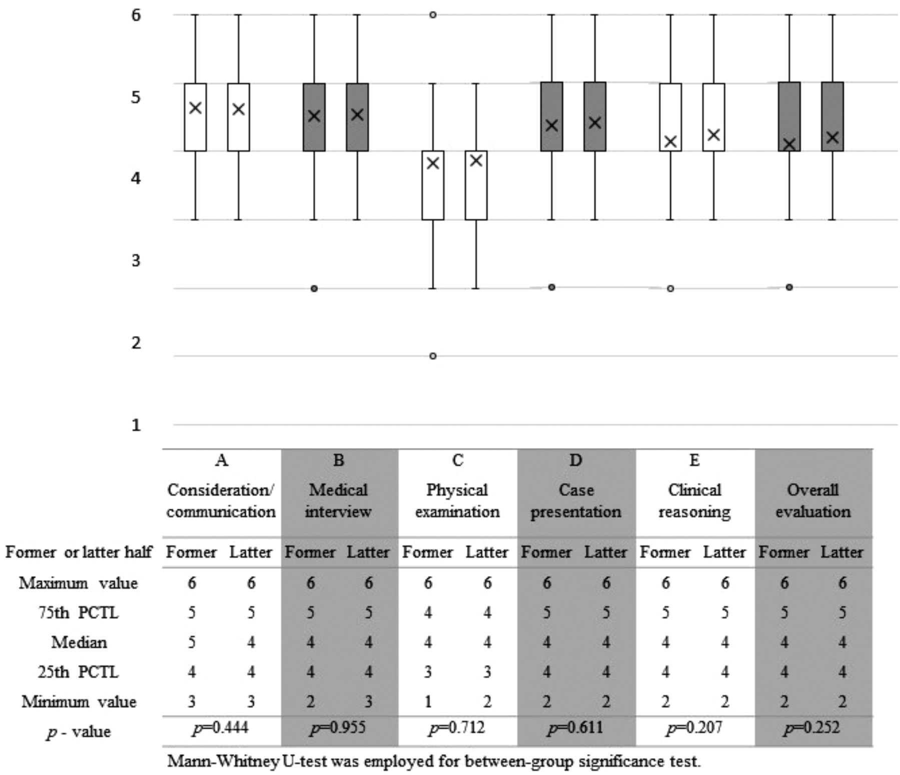
$p = 0.023$ ($\chi 2$ test)

Position-associated difference

The grade evaluation of each assessment item was calculated by the rater's position in all examinations (Fig. 5). A significant difference was noted in Items B (p = 0.011), C (p = 0.001), E (p = 0.005), and overall evaluation (p < 0.001). The evaluation by associate professors tended to be lower in three assessment items including the overall evaluation.

Discussion

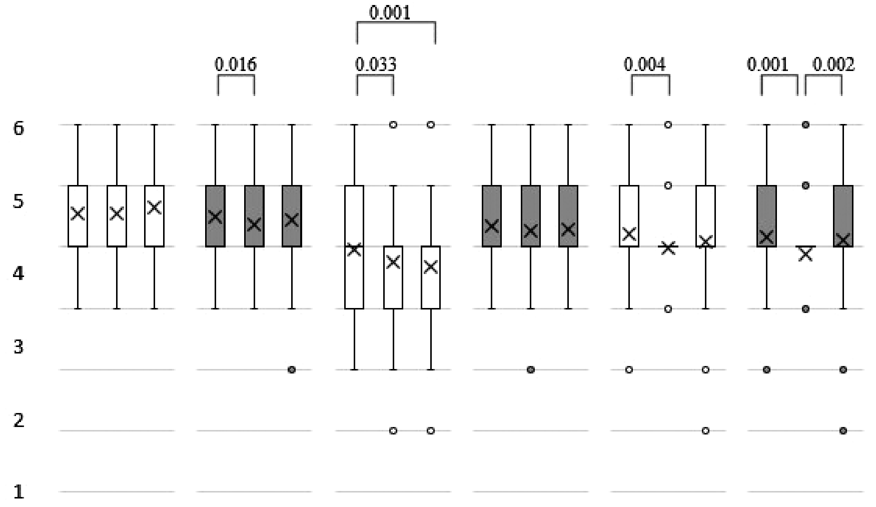
The survey was performed aiming at clarifying the reliability of evaluation of Post-CC OSCE initially performed officially. We demonstrated that approximately 60% of students scored between 3 and 4 for item C (physical examination), approximately 60% of students scored between 4 and 5 for the overall evaluation, and over 60% of students scored between 4 and 5 in the other 4 items. Overall, the



	A Consideration/ communication		B Medical interview		C Physical examination		D Case presentation		E Clinical reasoning		Overall evaluation	
Former or latter half	Former	Latter	Former	Latter	Former	Latter	Former	Latter	Former	Latter	Former	Latter
Maximum value	6	6	6	6	6	6	6	6	6	6	6	6
75th PCTL	5	5	5	5	4	4	5	5	5	5	5	5
Median	5	4	4	4	4	4	4	4	4	4	4	4
25th PCTL	4	4	4	4	3	3	4	4	4	4	4	4
Minimum value	3	3	2	3	1	2	2	2	2	2	2	2
p - value	p=0.444		p=0.955		p=0.712		p=0.611		p=0.207		p=0.252	

Mann-Whitney U-test was employed for between-group significance test.

Fig. 4 Median evaluation of each assessment item by examination-implemented time zone
Mann–Whitney U-test was employed for between-group significance test.



	A Consideration/ communication			B Medical interview			C Physical examination			D Case presentation			E Clinical reasoning			Overall evaluation		
	Professor	Associate professor	Lecturer	Professor	Associate professor	Lecturer	Professor	Associate professor	Lecturer	Professor	Associate professor	Lecturer	Professor	Associate professor	Lecturer	Professor	Associate professor	Lecturer
Maximum value	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
75th PCTL	5	5	5	5	5	5	4.5	4	4	5	5	5	5	4	5	5	4	5
Median	4	4	5	4	4	5	4	4	4	4	4	4	4	4	4	4	4	4
25th PCTL	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4	4	4	4
Minimum value	3	3	3	3	3	2	2	1	1	3	2	2	2	2	1	2	2	1
p - value	*p*=0.081			*p*=0.011			*p*=0.001			*p*=0.414			*p*=0.005			*P*<0.001		

Kruskal-Wallis test and Bonferroni correction was employed for significance test among 3 groups.

Fig. 5 Median evaluation of each assessment item by position
Kruskal–Wallis test and Bonferroni correction were employed for significance test among three groups.

scores of the study subjects were equivalent to those of the Post-CC OSCE performed across medical schools in Japan in 2020/2021.[12] Our findings also revealed that the time of the examination did not have a significant impact on the evaluation. When compared by the raters' position, we found that the scores for three assessment items as well as the overall evaluation were lower when assessed by associate professors. However, this is likely an alpha error as there are no logical explanations for this difference.

However, between the two raters within a station, κ coefficient of six assessment items including overall evaluation was Poor-Fair, the percent agreement was 41%-49%, being low, and disagreement of the evaluation score frequently extended over the passing line, 4 in "Item C: physical examination" and "Item E: clinical reasoning." Based on these findings, the evaluation by two raters within a station was deviated. Moreover, although no significant difference was noted in the overall evaluation score in each series, the overall evaluation by series was 90% (18/20) in series with a high evaluation and 36.8% (7/19) in series with a low evaluation, showing a large difference. In this section, measures to improve the reliability of the evaluation of Post-CC OSCE are considered.

The number of grades of rubric assessment

The percent agreement of the 6-grade evaluation between the 2 raters was at the 40% level, being low. Studies demonstrated that an assessment is more reliable if the number of assessment scale is reduced; specifically, a study found that κ coefficients of 0.5 or higher were achieved in more than half of cases when 2 raters performed 2- to 3-point assessment scales,[9] and another study showed that the percent agreement exceeded 80% in all tasks when a 6-point assessment scale was switched to a 3-point assessment scale.[13] It is important to adjust the assessment scale, particularly to define the scores for pass, borderline pass, and fail, and to ensure that the raters are well aware of these definitions to standardize the assessment criteria. However, although a reduction of assessment grades increases the percent agreement and simplifies assessment, this is a fake agreement, and it is hard to say that the reliability of the examination improved. At the same time, the reduction of assessment grades decreases information for assessment, and the data are very likely to be nonapplicable for various analyses. Since it is possible to apply conversion reducing the number of grades after assessment, the number of assessment grades should not be easily reduced as long as the burden on raters allows it.

Review of examination implementation system

The reproducibility between the two raters in this survey was a level of κ = 0.1-0.2, being low, and disagreement frequently extended over the passing line, 4, strongly influencing students, for which improvement is necessary. Hara et al.[6] mentioned increasing the number of raters to three or more as a measure to improve the reliability of OSCE. By increasing the number of raters to five or more within a station and determining the trimmed mean, reliability of evaluation may markedly increase, but considering the environment of medical schools including manpower in Japan, it is not realistic. About 10% of medical schools in Japan implemented Post-CC OSCE assessed by three raters,[12] and three raters are sufficient for investigation, but it requires 1.5 times the current manpower. Activities and assessment fields of university faculty members are divided into research, education, operation management, and contribution to society,[14] and clinical practice is added to these in medical school. Among these, the performance of education is difficult to quantitate and visualize. Considering that the assessment of faculty members is difficult because of this,[15] it is presumed that clinicians accounting for the majority of medical school faculty members cannot readily spare efforts sufficiently to the educational activity. Measures taken by the whole organization may be necessary to enable faculty members to spare efforts to the educational activity. In addition, it is necessary to investigate the introduction of assessment of SP in addition to faculty members from the viewpoint of 360-degree feedback.[5,7] However, in this case, there are many problems to be solved for its operation, such as an increase in the burden on SP and the possibility of the direction of negative emotion of students who failed to standardize patients.

Necessity of unifying interrater assessment criteria

OSCE is an assessment of practice, aiming at assessing attitude and skills not measurable by cognitive behavioral therapy. However, the assessment criteria in the rubric of performance assessment cannot be completely expressed by the criteria and matrix of the scale,[16] as interrater differences in the evaluation were clarified in this survey. Therefore, it is important to align the details of the assessment criteria unable to be documented between raters. For this alignment, a rater certification course was implemented, and the time to align the interrater assessment criteria between raters is set by the time of initiation of examination, but as far as analyzing the results of this survey, this is not sufficient. Devices for raters to secure time for alignment of the assessment criteria during working hours are necessary.

In addition, clearly, specifying anchors of assessment rubric as a tool for sharing the assessment criteria between raters is necessary. "Anchors" are defined as "Samples of work or performance used to set the specific performance standard for each level of a rubric" and are said to "contribute significantly to scoring reliability".[17] Miyawaki et al.[18] reported that the reliability of the evaluation was lower for the first two examinees than those of later examinees, suggesting that it is desirable to perform a mock OSCE immediately before the start of the actual OSCE. These findings suggest that the reliability of an assessment improves if the assessment criteria were shared using actual examples such as the mock OSCE and anchors. However, to prepare anchors, it is necessary to collect the performance of students assessed following the rubric in addition to their score data. Official implementation of Post-CC OSCE has just started in fiscal 2020 in Japan, and the preparation of anchors is a task in the future.

In this survey, the disagreement of evaluation directly linked to pass/fail of students was observed, and this is a big disadvantage for students. As described above, it may be necessary to secure a sufficient number of raters and minimize disagreement of evaluation by sharing the assessment criteria between raters. In addition, the construction of a system capable of reviewing evaluation may be necessary for cases in which disagreement very disadvantageous for students is observed on statistical confirmation of the evaluation after examination, as observed in this analysis.

Limitation of the study

This survey was performed aiming at the evaluation of the reliability of Post-CC OSCE officially implemented initially in a state unified by the Common Achievement Tests Organization, but the content and timing of implementation varied among universities in fiscal 2020 due to the influence of the COVID-19 pandemic.[4] Moreover, the number of tasks was reduced to half, three, in the implementation, reducing the sample size to half of that planned. It is necessary to perform a survey using test data collected from an increased number of tasks of Post-CC OSCE performed in a unified state nationwide after the following fiscal year. We were also unable to determine possible reasons for poor interrater agreement. Future studies should consider the addition of variables other than the raters' positions and time of the assessment to identify specific measures to help improve the interrater agreement.

Conclusions

1. The percent agreement was 41%-49%, and the κ coefficient was Poor-Fair, and it was low in six assessment items, including overall evaluation between the two raters within a station. Especially, disagreement of the evaluation frequently extended over the passing line, 4, in "Item C: physical examination" and "Item E: clinical reasoning."
2. Disagreement of evaluation directly linked to pass/fail of students was observed, and this is a big disadvantage for students. In addition to refining the implementation method and assessment method, it is important to statistically confirm the evaluation after examination and construct a system capable of reviewing the evaluation as needed.

Author's contribution: M.T. and N.H. designed the study.
M.T. and A.N. collected the data and analyzed the data.
N.H. supervised data collection and statistical analysis.
M.T. wrote the manuscript.

Ethics statement: This study was approved by the Ethics Committee of the Toho University School of Medicine (approval No. A20051_A19089).

To obtain consent from the students, an explanatory document was sent to the subjects by mail, and when a subject did not give consent, it was informed to the investigator. The same method was used to obtain consent from faculty members, who were raters, in which the mail to students was sent by an investigator who was not a faculty member of the medical school to minimize the pressure on students. Since the data contained students' records, the survey was initiated after a pass/fail judgment was made.

Conflicts of interest: None declared.

References

1) Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975; 1: 447-51.
2) Ban N, Tsuda T, Tasaka Y, Sasaki H, Kasai R, Wakunami M, et al. [Introducing the Objective Structured Clinical Examination to evaluate students' interviewing and physical examination skills]. Medical Education. 1994; 25 (6): 327-35. Japanese.
3) Ministry of Health, Labour and Welfare. Report from Medical Ethics Council (Physicians subcommittee) report: Communalization of common tests aiming at seamless training of physicians and legal positioning of so-called Student Doctor. 2020. https://www.mhlw.go.jp/content/10803000/000609623.pdf. Japanese.
4) Common Achievement Tests Organization (CATO). Crisis management policy for COVID-19 infection (1st report) [Internet]. 2020. http://www.cato.umin.jp/pdf/taro_200330-02.pdf (cited 2021 Dec. 12). Japanese.
5) Makuuchi A, Takemoto Y, Shimazaki I, Namikawa H, Kobayashi M, Kinuhata S, et al. Concurrences and differences between faculty member staff and standardized patients in the assessment of medical students in the Post-Clinical Clerkship Objective Structured Clinical Examination. Osaka City Med J. 2018; 64: 1-8.
6) Hara S, Tamai T, Ota K, Yamamoto Y, Nomura H. [Reliability assessment of 3-year Post-Clinical Clerkship OSCE using the generalizability theory]. Medical Education. 2020; 51 Suppl: 188. Japanese.
7) Murakami J, Takenama H, Horikoshi A, Sawada U, Sato M, Ohi H, et al. [Problems in the assessment of medical interviewing skills with Objective Structured Clinical Examinations: How can reasonable objectivity be ensured?]. Medical Education. 2001; 32 (4): 231-7. Japanese.
8) Suzuki E, Ito M, Aoyagi Y, Fuse I, Tanaka K, Naito M, et al. [Study of the suitability and reliability of assessments of initial Objective Structured Clinical Examinations at the Niigata University School of Medicine]. Medical Education. 2003; 34 (1): 37-44. Japanese.
9) Iwahori M, Ogawa M, Hirose S, Yoneda H, Sumitomo S, Muramatsu Y, et al. [Study on conformity between difference of inspectors of the OSCE results]. J. Gifu Dent. Soc. 2009; 35 (3): 160-6. Japanese.
10) Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159-74.
11) Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86: 420-8.
12) Common Achievement Tests Organization (CATO). Common achievement tests for medical and dental students. 19th ed. 2021. p. 161-2. http://www.cato.umin.jp/e-book/19/index.html#page=

1. Japanese.

13) Mikasa H, Akaike M, Terashima Y, Tani K, Takayama T, Fukui Y. Analysis of item-characteristic curve of common achievement test OSCE and inter-rater variation [Internet]. 2009. https://www.tokushima-u.ac.jp/fs/1/1/9/5/4/0/_/54-55D1.pdf (cited 2021 Dec. 12). Japanese.

14) Shimada T, Okui M, Hayashi T. [Progresses and problems of faculty member assessment system in Japanese universities]. Research on Academic Degrees and University Assessment. 2009; 10: 61-78. Japanese.

15) Kawabe T, Hano T, Sohma H, Suzuki K, Akaike M, Kobayashi N, et al. [Investigation of educational achievements of medical department faculty member members and healthcare providers using a rating form to evaluate medical education performance]. Medical Education. 2016; 47 (2): 77-88. Japanese.

16) Yamada Y, Mori T, Mouri M, Iwasaki C, Tanaka T. [Methodology concerning rubric assessment utilized for learning]. Kansai University Higher Education Research. 2015; 6: 21-30. Japanese.

17) Wiggins G, McTighe J. Understanding by design. 2nd ed. Alexandria: ASCD; 2005. p. 336.

18) Miyawaki S, Deguchi T, Murakami K, Fukunaga T, Kamioka H, Yoshida T, et al. [Reliability of evaluation immediately after initiation of explanatory OSCE]. JJDEA. 2007; 23 (3): 299-304. Japanese.