THE ASIA-PACIFIC SCHOLAR
MEDICAL AND HEALTH PROFESSIONS EDUCATION

ORIGINAL ARTICLE

Check for updates

Submitted: 28 July 2020
Accepted: 18 November 2020
Published online: 4 May, TAPS 2021, 6(2), 48-56
https://doi.org/10.29060/TAPS.2021-6-2/OA2367

Does objective structured clinical examination examiners' backgrounds influence the score agreement?

Oscar Gilang Purnajati[1], Rachmadya Nur Hidayah[2] & Gandes Retno Rahayu[2]

[1]Faculty of Medicine, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia; [2]Department of Medical Education, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract

Introduction: Objective Structured Clinical Examination (OSCE) examiners come from various backgrounds. This background variability may affect the way they score examinees. This study aimed to understand the effect of background variability influencing the examiners' score agreement in OSCE's procedural skill.

Methods: A mixed-methods study was conducted with explanatory sequential design. OSCE examiners (n=64) in the Faculty of Medicine Universitas Kristen Duta Wacana (FoM-UKDW) took part to assess two videos of Cardio-Pulmonary Resuscitation (CPR) competence to get their level of agreement by using Fleiss Kappa. One video portrayed CPR according to performance guideline, and the other portrayed CPR not according to performance guidelines. Primary survey, CPR procedure, and professional behaviour were assessed. To confirm the assessment results qualitatively, in-depth interviews were also conducted.

Results: Fifty-one examiners (79.7%) completed the assessment forms. From 18 background categories, there was a good agreement (>60%) in: Primary survey (4 groups), CPR procedure (15 groups), and professional behaviour (7 groups). In-depth interviews revealed several personal factors involved in scoring decisions: 1) Examiners use different references in assessing the skills; 2) Examiners use different ways in weighting competence; 3) The first impression might affect the examiners' decision; and 4) Clinical practice experience drives examiners to establish a personal standard.

Conclusion: This study identifies several factors of examiner background that allow better agreement of procedural section (CPR procedure) with specific assessment guidelines. We should address personal factors affecting scoring decisions found in this study in preparing faculty members as OSCE examiners.

Keywords: *OSCE Score, Background Variability, Agreement, Personal Factor*

Practice Highlights

- The examiners' background variability influences the OSCE scoring agreement results.
- The reason for assessment inaccuracy remains unclear regarding the score agreement.
- The absence of assessment instruments that could provide a loophole for examiners to improvise.
- Personal factors affecting scoring decisions found in this study should be addressed in preparing OSCE examiners.

I. INTRODUCTION

To assess medical students' competencies in a variety of skills, most medical schools in Indonesia implement the Objective Structured Clinical Examination (OSCE) both as a clinical skills examination at the undergraduate stage and as a national exit exam (Rahayu et al., 2016; Suhoyo et al., 2016). Most OSCE stations test both communication domains and specific clinical skills that will be assessed based on rubrics and scoring checklists which relies on examiners' observations (Setyonugroho et al., 2015). The OSCE has a challenge in its complexity to standardise the scores, which are very depend on OSCE examiners' perceptions (Pell et al., 2010). In a well-designed OSCE the examinees performance should only influence the examinees' score, with minimal effects from other sources of variance (Khan et al.,

2013). Research showed that there are influences of examiner's background variability on OSCE results although they have been asked to standardise their behaviour (Pell et al., 2010) The decision and behaviour of OSCE examiners will affect the quality of assessment, including making a pass or fail decision, considering the complexity of knowledge, skill, and attitude in medical education (Colbert-Getz et al., 2017; Fuller et al., 2017).

Examiners' observations also rely on their clinical practice experience, OSCE examining experience, and gender conformity (Mortsiefer et al., 2017). Even in OSCE that is held in the most standard conditions, the examiner factor has the biggest role in scoring inaccurately (Mortsiefer et al., 2017). However, the reason for this inaccuracy remains unclear since there are concerns regarding the scoring agreement of examiners in OSCE and how the result might be affected by this issue. There is a need to consider the influence of examiners' background variability (gender, educational level, clinical practice experiences, length of clinical practice experiences, OSCE experience, and OSCE training experience) when preparing teachers as OSCE examiners. This study aimed to understand background variability as a factor influencing examiners' scoring agreement in assessing students' performance in procedural skill, as the first step of faculty development program to ensure the standard quality for examiners.

II. METHODS

A. Study Design

This mixed-method study used a sequential explanatory design. This mixed-method approach is expected to provide more comprehensive results and better understanding than using a separated method (Creswell & Clark, 2018).

This study comprised of 2 sequential phases of data collection and analysis (QUANTITATIVE: qualitative) using sequential design. First, quantitative data were collected as a cross-sectional study of the examiners' strength of agreement using Fleiss Kappa while assessing the clinical skill performance recorded in the 2 videos: one video portrayed CPR according to performance guideline and the other portrayed CPR not according to performance guideline. We used these 2 videos in order to portray more comprehensively how the consistency of OSCE examiner agreement both on good and poor clinical skill performance.

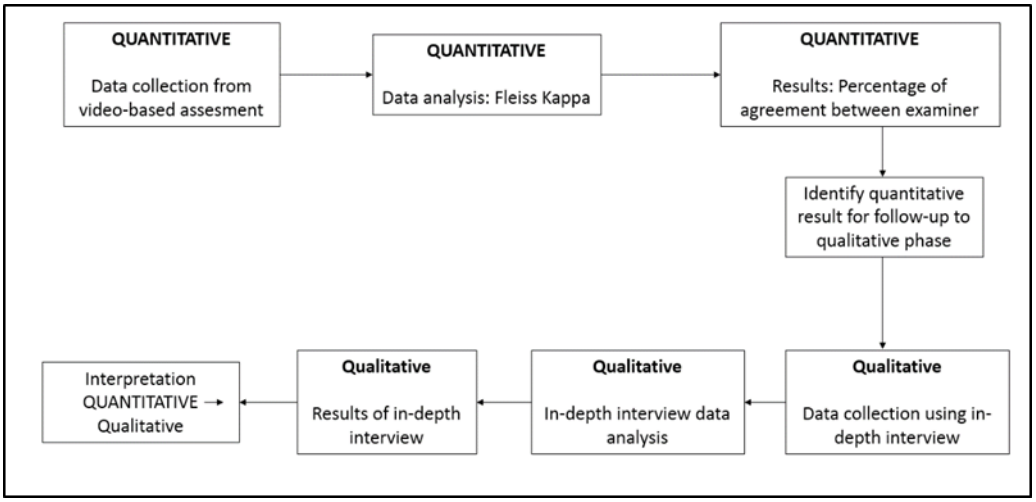

Figure 1 Mixed method explanatory design

In the second phase, in-depth interviews were used to complement the quantitative results to gain more information and a detailed confirmation about how the scores were decided (Stalmeijer et al., 2014). In this stage of study, researchers explored and explained the examiners' OSCE experiences and behaviour when they give a score on a clinical skill examination and the influences on their scoring regarding their backgrounds.

B. Materials and/or Subjects

The strength of agreement of the videos' score came from 64 OSCE examiners FoM UKDW. Mortsiefer et al., (2017), explained that more subjects are better when investigate examiner characteristics associated with inter-examiner reliability (Mortsiefer et al., 2017). In the second phase, in-depth interviews were conducted with 6 examiners of FoM UKDW, selected by purposive sampling regarding their scores and how they represented their own unique background (Table 1).

Researcher (OGP) provided all the participants with written information about this research and addressed ethical issues in an informed consent form. Researcher ensured participants understand the research protocol and clarified any questions regarding this study. Participants who agreed to take part, sign the informed consent form prior to the data collection.

We held interviews in FoM UKDW with maximum 30 minutes of duration each interview. The inclusion criteria for examiners who were selected for this study were involved as full-time faculty members, had over 4 times OSCE examination experience, and had done OSCE examiner training, expecting that they had enough interaction with other faculty members and had influences from medical doctor education (Park et al., 2015). The exclusion criteria were participant did not answer the research invitation and did not fill the assessment form completely. Main researcher (OGP) conducted the interview. Main researcher was a male, student of Master of Health Profession Education Universitas Gadjah Mada, and the staff of FoM UKDW.

C. Statistics

1)Quantitative data analysis: We grouped examiners into 18 groups based on their background which were gender, educational level, clinical practice experiences, length of clinical practice experiences, OSCE experience, and OSCE training experience as shown in Table 1. We analysed all gathered data using IBM SPSS Statistics 25 and Microsoft Office Excel 365 (IBM Corp., Chicago). We presented quantitative data as a strength of agreement in percentage. The strength of agreement was calculated using Fleiss Kappa to determine the agreement between each group of each examiner background on whether CPR performances (primary survey, CPR Procedure, and professional behaviour), that portrayed in those 2 videos, were exhibiting score either "0", "1", "2", or "3" based on the assessment guideline and rubric's criteria (Purnajati, 2020). Based on recent research, agreement above 60% was considered as a substantial and adequate agreement (Stoyan et al., 2017; Vanbelle, 2019).

2) Qualitative data analysis: In-depth interviews were analysed using thematic analysis. We prepared a structured list of questions. It consisted of one key question: What was your experience in scoring the OSCE? The other additional questions evaluated the experiences of examiners in OSCE scoring including: the use of other references, differences in assessment weighting, use of own decision, clinical practice experience affecting the decision, and gender related decision making. Next, the collected data resulting from in-depth interviews were recorded using audio file recorder, read, and categorised into themes whenever they were related. The transcripts and identified themes were then given to an external coder in this study. This step was followed by our agreement for each theme. There was no repeated interview.

III. RESULTS

A. Quantitative Data Result

We deposited both quantitative and qualitative data in an online repository (Purnajati, 2020). The study participants in this quantitative phase were 64 OSCE examiners who are full-time faculty members. Twelve participants were excluded because did not fulfil the inclusion criteria. Fifty-one (79.7%) examiners who returned the completed assessment form are described below in Table 1.

Quantitative Phase Participant		
Background	Groups	Number of Participant (*N=51*)
Gender	Male	22 (43%)
	Female	29 (57%)
Education	Bachelor undergraduate	19 (37%)
	Master's degree	16 (31%)
	Doctoral degree	3 (6%)
	Specialist doctor	13 (25%)
Clinical Practice Experience	General practitioner	28 (55%)
	Specialist	14 (27%)
	No clinical practice	9 (18%)
Duration of clinical practice experience	< 2 years	9 (18%)
	2-5 years	17 (33%)
	>5 years	25 (49%)
OSCE experience	< 2 years	9 (18%)
	2-5 years	24 (47%)
	>5 years	18 (35%)

OSCE examiner training	< 3 times	21 (41%)
	3-5 times	17 (33%)
	>5 times	13 (25%)

Qualitative Phase Participants.

ID	Score A[a]	Score B[b]	Gender	Education	Clinical Practice Experience	Clinical Practice Duration	OSCE Experience	OSCE Training Experience
23	48.15	33.33	Male	Master	GP	2-5 years	>5 years	3-5 times
28	100.00	44.44	Female	Doctoral	No Practice	>5 years	2-5 years	< 3 times
11	74.07	74.07	Male	Specialist	Specialist Doctor	>5 years	>5 years	>5 times
24	74.07	11.11	Male	Bachelor	GP	< 2 years	< 2 years	3-5 times
26	100.00	33.33	Female	Specialist	Specialist Doctor	>5 years	2-5 years	< 3 times
35	100.00	0.00	Female	Master	GP	>5 years	>5 years	>5 times

[a] Video portrayed CPR according to performance guideline. [b] Video portrayed CPR not according to performance guidelines

Table 1. Descriptive characteristics of participants

The assessment rubric was divided into three main competencies: (1) primary survey, (2) CPR procedure, and (3) professional behaviour. The results showed overall agreement on each main competency based on each examiners' background variability by using Fleiss Kappa. The percentage of agreement is shown in Figure 2, 3, and 4.

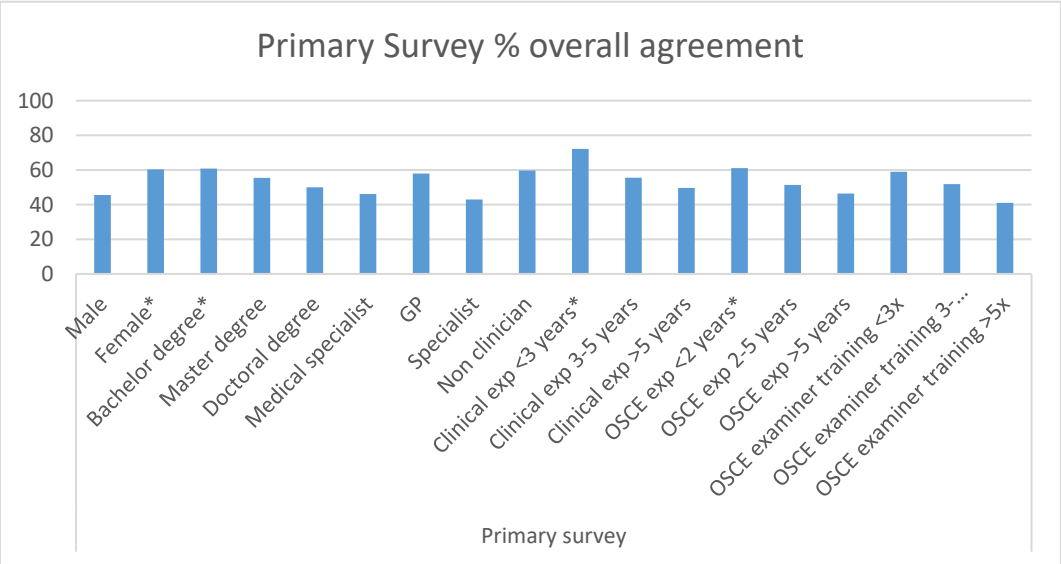

Figure 2. Primary Survey percentage of overall agreement (n = 51). Agreement above 60% (*) is considered as a substantial and adequate agreement

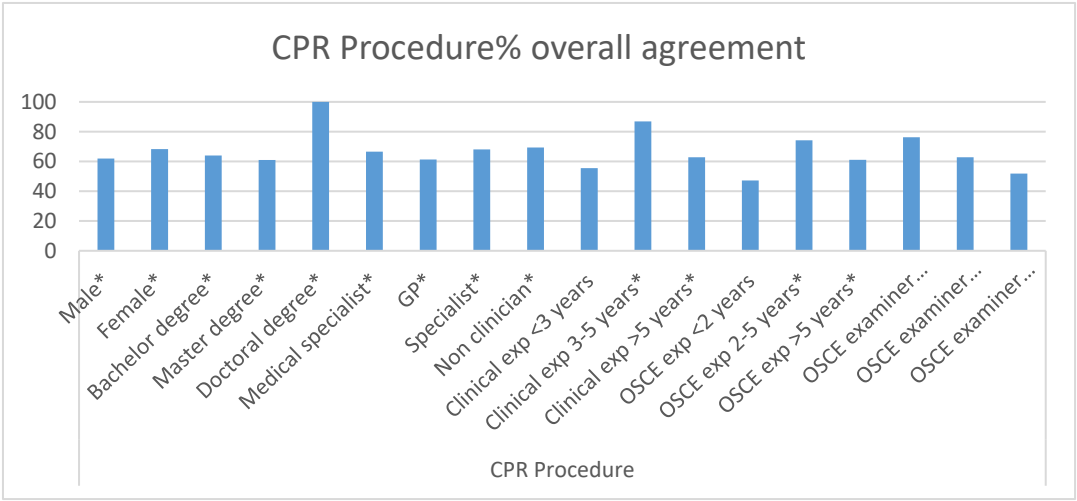

Figure 3. CPR Procedure percentage of overall agreement (n=51). Agreement above 60% (*) is considered as a substantial and adequate agreement

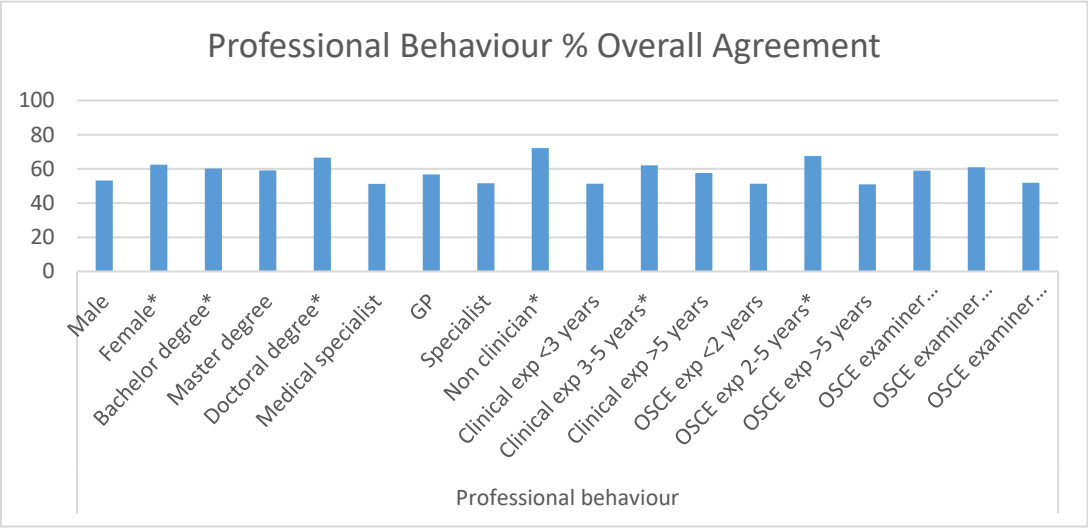

Figure 4. Professional Behaviour percentage of overall agreement (n=51). Agreement above 60% (*) is considered as a substantial

After completing the CPR competency assessment, all examiners' background characteristics met a cutoff of approval above 60% in assessing CPR procedure except for examiners with clinical practice experience <3 years, OSCE testing experience <2 years, and OSCE examiner training> 5 years (Figure 3). This finding showed a good strength of agreement in assessing CPR procedure regardless of examiners' background. However, there were many instances where the cut-off point of 60% was not achieved in the aspects of primary surveys and professional behaviour (Figure 2 and 4), which showed fair strength of agreement between examiners when they examined these competencies.

B. Qualitative Data Results

Two theme categories were determined: (1) OSCE experience and (2) specific behaviour in OSCE. The first theme contains of 3 sub-themes: (1) student performance, (2) examiner background effect, and (3) using assessment instrument. The second theme consists of 5 sub-themes: (1) use of assessment references, (2) score weighting, (3) personal inferences, (4) clinical experience, and (5) gender conformity.

Theme 1: Examiners argued that they understand the difference in student performance in performing clinical skills and can distinguish from the coherent skills performed by students according to checklist.

"Very easy in giving an assessment, because everything is in accordance with the assessment rubric"

(ID 35)

"The plot is clear, well organised"

(ID 26)

"You can compare the inadequacies; it is enough to be compared"

(ID 11)

"The 2 different students are quite striking, so in my opinion it is not too difficult"

(ID 28)

Nevertheless, some examiners had difficulty to distinguish student performance when only used a checklist. Examiner background did not affect their way in scoring clinical skills performance, but some background may have the potential to affect their scoring, such as clinical practice experience.

"I am trying to avoid personal interpretations, as much as possible, but of course that cannot be 100 percent. In my opinion, the assessment rubric still gives room for subjectivity"

(ID 28)

In this research, it seemed easy for examiners to understand the assessment instrument when giving score to those 2 videos and their understanding were good.

Theme 2: Interviews revealed that: 1) Examiners use other references such as their clinical experience in assessing the skills;

"If the assessment guideline is unclear, the students are also unclear, yes I will improvise. Or when the assessment guideline is clear and the students are unclear which criteria are included, yes I will improvise"

(ID 35)

"Maybe yes, because once again the template at the beginning is not very clear"

(ID 23)

2) Examiners use different ways in giving weight of competence, for example, procedural steps are considered more important than primary survey;

"For those that I feel have a small weight because the instructions are also short, so I don't have to look carefully"

(ID 24)

"When I feel that competence is not important, it does not get my emphasis, the more emergency that will get more attention."

(ID 28)

3) The first impression of examinees might affect their decision in scoring their performance;

"That first impression will affect me in giving value. I will be more critical. I see more, pay more attention to the small things they do"

(ID 24)

4) Clinical practice experience drives examiners to establish a personal standard on how a doctor should be;

"*Clinical experience when practice is one of the judgments"*

(ID 24)

"The reference is just my instinct because it has been running as a doctor after all these years. Yes, I use my previous knowledge"

(ID 26)

And 5) Gender of examinees does not affect their decision, while their professionalism (e.g. showing respect to patients) will surely affect their decision.

"I pay more attention especially to politeness and professional behaviour"

(ID 24)

"Students of any gender still have the same standard of evaluation, a score of professionalism which is more influential"

(ID 23)

IV. DISCUSSION

Examiners' agreement in this study was high in assessing the CPR procedure, which has a fixed and specific procedure in almost all groups of examiners. These results are consistent and can be explained by results from previous studies, which show that assessment with specific cases will provide high inter-examiner agreement (Erdogan et al., 2016). The differences in the examiner's background will not have much influence on their agreement in giving an assessment in a specific case. This was supported by the opinions of examiners in the in-depth interviews who stated that in the CPR assessment procedure, assessment instruments are clear, easy to understand, with clear procedure flow, and performance that is easily distinguished, which made it easier for examiners to be able to distinguish student performance. A specific assessment instrument that could not provide a loophole for examiners to improvise assessment, made the opportunity for examiners to portray their subjectivity was minimised. This simplicity could lead to high agreement among examiners in specific competencies as shown in this study and based on clear evidence can increase the reliability of the assessment (Daniels et al., 2014) .

In this study, it was found in the primary survey assessment and professional behaviour which has an assessment guide that is not as specific as the CPR procedure, the percentage of agreement between examiner groups was lower, with only a few of them reaching 60% of agreement. This difference happened for reasons confirmed in the in-depth interviews which raised the issue that although the examiners tried to minimise their subjectivity in assessing, but it was said that there were still gaps in the assessment guide that still gives room for subjectivity. There are also examiners who were dissatisfied with the checklist, so they used their personal decisions in evaluating students.

According to a recent study, this could be due to the lack of specific instructions in the general assessment guidelines which will result in lower inter-examiner reliability compared to the use of more specific assessment guidelines (Mortsiefer et al., 2017). In the primary survey section and professional behaviour, there were also aspects of communication that were judged to be more susceptible to bias than physical examination skills because physical examination is more well-documented, clear instructions, and more widely accepted by examiners (Chong et al., 2018) The validity and reliability of a clinical skills assessment depend on factors including how the student's performance on the exam, the character of the population, the environment, and even the assessment instrument itself can affect how examiners carry out the assessment (Brink & Louw, 2012). These phenomena were seen in the in-depth interviews which revealed that there were certain moments namely when the student being tested does not match the expectations written in the assessment guide and when the assessment guide is not clear so that it still gives room for subjectivity examiner. In addition, in the

in-depth interviews the results also revealed that the examiners differentiated their attention on certain competencies with certain criteria such as the length of information in the assessment rubric, so that competencies that were considered not important did not get as much attention.

This finding may be in line with previous research which stated that constructs and conceptual definitions in this category that still provide a gap in the subjectivity of examiners cause shifting attention focus and weighting of their judgments to be different so that there are differences in important aspects between examiners (Schierenbeck & Murphy, 2018; Yeates et al., 2013). The difference in these important aspects can bring examiners to reorganise competency weights so that simpler and easier competencies (in this case those that have clearer and more detailed assessment guidelines) will be done first, and more complex ones (in this case, guides that have lower rigidity ratings) will be assessed later with the possibility of using more narratives (Chahine et al., 2015). This reorganisation can reflect how the examiners' decision, allowing them to direct their attention to the more important aspects as the testers revealed in in-depth interviews with this research.

The personal factor, such as assessment references is a potential variability of the assessment conducted by the examiner. Examiners are trained and understand the use of assessment instruments, but produce varying assessments because they do not apply assessment criteria appropriately, but use personal best practice, use other test participants better as benchmarks, use patient outcomes (e.g. correct diagnosis, do patients understand, etc.), and use themselves as a comparison (Gingerich et al., 2014; Kogan et al., 2011; Yeates et al., 2013).

Another personal factors, including first impressions, can occur spontaneously unconsciously and can be a source of difference in judgment between examiners (Gingerich et al., 2011). First impressions based on observers' observations have the same decisions and influences as social interactions, so it makes sense that first impressions are able to influence judgments, can be accurate and have a relationship with the final assessment results, but do not occur in examiners in general (Wood, 2014; Wood et al., 2017).

In providing assessments, there are gaps for examiners to give different competency weights to other examiners. Providing assessments based on targets that differ from competency standards and comparisons with the performance of other examinees will make the examiners recalibrate their own weighting and this is an explanation why there are variations in assessment and differences in the important points of the examinees' performance among examiners (Gingerich et al., 2018; Yeates et al., 2015; Yeates et al., 2013).

The variability of personal factors between examiners can be conceptualise more as a different emphasis on building doctor-patient relationships and / or certain medical expertise rather than variations in the examiner's background itself. The examiners' own understanding can be conceptualized as a combination of whether what the examinees do is good enough and whether what they do is enough to build a doctor-patient relationship.

This research had some limitations such as it only used specific cases (i.e., CPR) to minimise the bias of the assessment instrument so that it would reveal more bias in the examiners themselves. In more complicated cases such as communication skills and clinical reasoning it is also necessary to provide a more complete picture of how the examiners' scores agree in other cases. Generalization also became a limitation in this study because it only involved examiners from one medical education institution, however the study participants sufficiently described the variability of the examiner's background.

V. CONCLUSION

This study identifies several factors of examiner background variability that influence examiners' judgment in terms of inter-examiner agreement. Female examiners, bachelor education, less OSCE experience, and non-clinician examiners allow better agreement of procedural section (CPR procedure) with specific assessment guidelines. Cases that have unspecified assessment guidelines in this research, primary survey and professional behaviour, have lower agreement among examiners and must be examined deeper. We should note that personal factors of OSCE examiners can influence assessment discrepancies. However, the reasons for using these personal factors in scoring OSCE performance might be affected by unknown biases that require further research. Therefore, to improve clinical skills assessment such as OSCE for undergraduate medical programme, we must address personal factors affecting scoring decisions found in this study in preparing faculty members as OSCE examiners.

Notes on Contributors

Oscar Gilang Purnajati, MD was student of Master of Health Professions Education Study Program, Faculty of Medicine, Universitas Gadjah Mada, Indonesia. He concepted the research, reviewed the literature, designed the study, acquisited funding, conducted interviews, analysed quantitative data and transcripts, and wrote the manuscript.

Rachmadya Nur Hidayah, MD., M.Sc., Ph.D is lecturer of Department of Medical Education, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia. She supervised author Oscar Gilang Purnajati, developed the concepted framework for the study, critically analysed the data, cured the data, and reviewed the final manuscript.

Prof. Gandes Retno Rahayu, MD., M.Med.Ed, Ph.D is professor at the Department of Medical Education, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia. She supervised author Oscar Gilang Purnajati, advised the design of the study, critically analysed the data, gave critical feedback to the conducted interviews, reviewed the final manuscript.

All the authors have read and approved the final manuscript.

Ethical Approval

This study was approved by Health Research Ethics Committee Faculty of Medicine Universitas Kristen Duta Wacana (Reference No.1068/C.16/FK/2019).

Data Availability

All data were deposited in an online repository. The data is available at Open Science Framework with DOI: https://doi.org/10.17605/OSF.IO/RDP65

Acknowledgements

The author would like to thank Hikmawati Nurrokhmanti, MD, M.Sc for helping with the process of coding the in-depth interview transcripts. The author also would like to thank the staffs of Faculty of Medicine, Universitas Kristen Duta Wacana for supporting the research.

Funding Statement

This work was supported by the Universitas Kristen Duta Wacana (No. 075/B.03/UKDW/2018) as a part of study scholarship.

Declaration of Interest

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

Abbreviations and specific symbols

OSCE: Objective Structured Clinical Examination.

References

Brink, Y., & Louw, Q. A. (2012). Clinical instruments: Reliability and validity critical appraisal. *Journal of Evaluation in Clinical Practice*,*18*(6), 1126-1132. https://doi.org/10.1111/j.1365-2753.2011.01707.x

Chahine, S., Holmes, B., & Kowalewski, Z. (2015). In the minds of OSCE examiners: Uncovering hidden assumptions. *Advances in Health Sciences Education : Theory and Practice*, *21*(3), 609–625. https://doi.org/10.1007/s10459-015-9655-4

Chong, L., Taylor, S., Haywood, M., Adelstein, B.-A., & Shulruf, B. (2018). Examiner seniority and experience are associated with bias when scoring communication, but not examination, skills in objective structured clinical examinations in Australia. *Journal of Educational Evaluation for Health Professions*, *15*(17). https://doi.org/10.3352/jeehp.2018.15.17

Colbert-Getz, J. M., Ryan, M., Hennessey, E., Lindeman, B., Pitts, B., Rutherford, K. A., Schwengel, D., Sozio, S. M., George, J., & Jung, J. (2017). Measuring assessment quality with an assessment utility rubric for medical education. *MedEdPORTAL : The Journal of Teaching and Learning Resources*, *13*, 1-5. https://doi.org/10.15766/mep_2374-8265.10588

Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed method research*. SAGE Publications, Inc.

Daniels, V. J., Bordage, G., Gierl, M. J., & Yudkowsky, R. (2014). Effect of clinically discriminating, evidence-based checklist items on the reliability of scores from an internal medicine residency OSCE. *Advances in Health Sciences Education : Theory and Practice*, *19*(4), 497-506. https://doi.org/10.1007/s10459-013-9482-4

Erdogan, A., Dong, Y., Chen, X., Schmickl, C., Berrios, R. A. S., Arguello, L. Y. G., Kashyap, R., Kilickaya, O., Pickering, B., Gajic, O., & O'Horo, J. C. (2016). Development and validation of clinical performance assessment in simulated medical emergencies: An observational study. *BMC Emergency Medicine*, *16*, 4. https://doi.org/10.1186/s12873-015-0066-x

Fuller, R., Homer, M., Pell, G., & Hallam, J. (2017). Managing extremes of assessor judgment within the OSCE. *Medical Teacher*, *39*(1), 58-66. https://doi.org/ 10.1080/0142159X.2016.1230189

Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education*, *48*(11), 1055–1068. https://doi.org /10.1111/medu.12546

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Academic Medicine : Journal of the Association of American Medical Colleges*, *86*(10), S1-S7. https://doi.org/10.1097/ACM.0b013e31822a6cf8

Gingerich, A., Schokking, E., & Yeates, P. (2018). Comparatively salient: Examining the influence of preceding performances on assessors' focus and interpretations in written assessment comments. *Advances in Health Sciences Education: Theory and Practice*, *23*(5), 937-959. https://doi.org /10.1007/s10459-018-9841-2

Khan, K. Z., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher*, *35*(9), 1437-1446. https://doi.org/ 10.3109/0142159X.2013.818634

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: A conceptual model. *Medical Education*, *45*(10), 1048-1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x

Mortsiefer, A., Karger, A., Rotthoff, T., Raski, B., & Pentzek, M. (2017). Examiner characteristics and interrater reliability in a communication OSCE. *Patient Education and Counseling*, *100*(6), 1230-1234. https://doi.org/ 10.1016/j.pec.2017.01.013

Park, S. E., Kim, A., Kristiansen, J., & Karimbux, N. Y. (2015). The Influence of Examiner Type on Dental Students' OSCE Scores. *Journal of Dental Education*, *79*(1), 89-94.

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics – AMEE guide no. 49. *Medical Teacher*, *32*(10), 802-811. https://doi.org/10.3109/0142159X.2010.507716

Purnajati, O. G. (2020). *Does objective structured clinical examination examiners' backgrounds influence the score agreement? [Data set]*. Open Science Framework. https://doi.org/ 10.17605/OSF.IO/RDP65

Rahayu, G. R., Suhoyo, Y., Nurhidayah, R., Hasdianda, M. A., Dewi, S. P., Chaniago, Y., Wikaningrum, R., Hariyanto, T., Wonodirekso, S., & Achmad, T. (2016). Large-scale multi-site OSCEs for national competency examination of medical doctors in Indonesia. *Medical Teacher*, *38*(8), 801-807. https://doi.org/10.3109/0142159X.2015.1078890

Schierenbeck, M. W., & Murphy, J. A. (2018). Interrater reliability and usability of a nurse anesthesia clinical evaluation instrument. *Journal of Nursing Education*, *57*(7), 446-449. https://doi.org/10.3928/01484834-20180618-12

Setyonugroho, W., Kennedy, K. M., & Kropmans, T. J. B. (2015). Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Education and Counseling*, *98*(12), 1482-1491. https://doi.org/ 10.1016/j.pec.2015.06.004

Stalmeijer, R. E., McNaughton, N., & Van Mook, W. N. (2014). Using focus groups in medical education research: AMEE Guide No. 91. *Medical Teacher*, *36*(11), 923-939. https://doi.org/10.3109/0142159X.2014.917165

Stoyan, D., Pommerening, A., Hummel, M., & Kopp-Schneider, A. (2017). Multiple-rater kappas for binary data: Models and interpretation. *Biometrical Journal*, *60*(5), 381-394. https://doi.org/ 10.1002/bimj.201600267

Suhoyo, Y., Rahayu, G. R., & Cahyani, N. (2016). A national collaboration to improve OSCE delivery. *Medical Education*, *50*(11), 1150–1151. https://doi.org/ 10.1111/medu.13189

Vanbelle, S. (2019). Asymptotic variability of (multilevel) multirater kappa coefficients. *Statistical Methods in Medical Research*, *28*(10-11), 3012-3026. https://doi.org /10.1177/0962280218794733

Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education : Theory and Practice*, *19*(3), 409-427. https://doi.org/ 10.1007/s10459-013-9453-9

Wood, T. J., Chan, J., Humphrey-Murto, S., Pugh, D., & Touchie, C. (2017). The influence of first impressions on subsequent ratings within an OSCE station. *Advances in Health Sciences Education : Theory and Practice*, *22*(4), 969-983. https://doi.org/10.1007/s10459-016-9736-z

Yeates, P., Moreau, M., & Eva, K. (2015). Are examiners' judgments in osce-style assessments influenced by contrast effects? *Academic Medicine : Journal of the Association of American Medical Colleges*, *90*(7), 975-980. https://doi.org/10.1097/ACM.0000000000000650

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education : Theory and Practice*, *18*(3), 325-341. https://doi.org/10.1007/s10459-012-9372-1

*Oscar Gilang Purnajati
Faculty of Medicine,
Universitas Kristen Duta Wacana,
Jl. Dr. Wahidin Sudirohusodo No. 5-25.
Yogyakarta City,
Special Region of Yogyakarta
55224, Indonesia.
Tel: +62-274-563929
Email: oscargilang@staff.ukdw.ac.id