Feasibility of Peer Assessment in Evaluating Medical Students' Professional Behaviour in the Early Years of Their Medical Journey

Dujeepa D. Samarasekera, Lee Shuh Shing, Yeo Su Ping, and Denise Goh

Abstract Peer evaluation/assessment is used by many programmes to assess medical students' professional behaviour. However, there are doubt about peers' ability and consistency to evaluate their peers. We conducted a pilot study to explore the method's reliability: a quantitative study, involving Phase I and II medical students from the National University of Singapore, was conducted utilising repeated peer assessments in a small group teaching programme, Collaborative Learning Cases (CLC). A 5-question online form on a 9-point Likert scale was used. Descriptive and Gwen's agreement coefficient analysis were done using SPSS. 52 Phase I and 54 Phase II students participated. Average scores for most questions for Phase I and II students were higher at the last session as compared to the first session. In terms of combined inter-reliability, more "perfect agreement" was observed by the mid and last sessions. Results suggest that peer assessment could be a reliable tool in assessing peers' professional behaviour. However, for this to be effective, the students must be given clear guidelines.

Keywords Pilot studies · Educational measurement · Program evaluation · Peer assessment · Professionalism

D. D. Samarasekera (✉) · L. S. Shing · Y. S. Ping
Centre for Medical Education, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
e-mail: dujeepa@nus.edu.sg

D. Goh
Department of Pediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Department of Pediatrics, National University Healthcare System, Singapore, Singapore

M. Claramita et al. (eds.), *Character Building and Competence Development in Medical and Health Professions Education*, Springer Proceedings in Humanities and Social Sciences, https://doi.org/10.1007/978-981-99-4573-3_11

1 Introduction

Background/rationale

Professionalism embodies a vital skill that medical practitioners are expected to emulate to the highest standards. The American Board of Medical Specialties (ABMS) defines medical professionalism as "a belief system about how best to organize and deliver health care, which calls on group members to jointly declare ("profess") what the public and individual patients can expect regarding shared competency standards and ethical values and to implement trustworthy means to ensure that all medical professionals live up to these promises" [1].

Many medical schools nurture their undergraduates with professionalism early by incorporating this in the curriculum and assessing students regularly. Many teaching modalities (e.g. role modeling, didactic lectures, reflection, interactive methods, etc.) and evaluation methods for professionalism exist [2]. Due to the complexity of professionalism, multiple assessment tools are usually used including multi-source feedback using 360-degree reviews, critical incident reports and patient feedback [2]. These assessment tools are incorporated during their clinical year instead of early year in their medical training. As professional behaviors are important since day one in medical school, peer assessment has emerged as a possible alternative method for providing formative peer feedback to students regarding their professional behaviour [3–6].

Peer assessment is commonly used to assess student professionalism as it is simple to use and also has some advantages [7]. For instance, it has been demonstrated that peer feedback during collaborative learning settings in undergraduate medical education denote dependable methods of assessment for professionalism, and serves to support the advancing of professional behaviour over time [7, 8]. Additionally, studies suggest that peer assessment offers advantages in terms of shedding light on non-cognitive qualities such as team contribution and personal attributes, and these are likely to be correlated with professional and interpersonal skills [9]. Use of peer assessment data may also create a culture of learning whereby students feel that their experiences and inputs are valued [9]. As peers are the closest person within a class that they collaborate with and are able to observe one another regularly over a wide range of circumstances, peer assessment may also provide information regarding student behaviour that is not measured by other traditional evaluation methods [3, 10–14].

Despite the benefits of peer assessment, some research studies have suggested that peer assessment represents a less reliable means of assessment. This is because the quality of peer assessment can be influenced by multiple factors, such as the reliability of the assessment, the interaction between peers, the stakes of the assessment and the assumption of equivalence between the evaluations of each (student) colleague or peer [15]. It has been observed that students are reluctant to take part in peer assessment especially in a face-saving and conflict avoidance culture among

Asian students [16, 17]. As a result, very few peer assessment systems have been implemented and published in Asian setting.

Given that active participation and accurate, appropriate and meaningful peer assessment may be constrained by fear of mistakes, politeness norms, and the belief that peer feedback lacks credibility compared with teacher feedback,, we conducted a pilot study to explore whether repeated peer assessment by preclinical students help in improving professionalism, and whether the method is realiable as a formative evaluation tool.

Objectives:

This pilot study seeked to explore whether utilising repeated peer assessment in a small group teaching programme is a reliable evaluation tool and if it helps in improving professionalism.

2 Methods

Ethics approval was obtained from the Institutional Review Board, National University of Singapore (Reference number: S19-342). Informed consent was taken from the student participants.

Study design

This was a quantitative cross-sectional pilot study involving a subset of students from each cohort.

Setting

NUS Medicine introduced Collaborative Learning Cases (CLCs) as a structured small group teaching program that is part of the Phase I and II curriculum. The programme occurs through a series of small group sessions, instructed by clinician/biomedical science educators paired tutors, that review prototypical clinical cases. There are around 8–10 sessions per academic year, spaced out over 5 months. Each session covers a different topic e.g. allergy, anaemia (thalassaemia), sudden breathlessness, Parkinson's, and joint pain.

The cohort, of around 300 students per phase, was separated into 100 groups, each comprising of 5–7 students. During the CLC group sessions, students are engaged individually in proposing and discussing different approaches through real time live interactions within a group [18].

Participants

Ten student groups within each cohort were randomly selected for the pilot and invited to participate which took place from August to December 2020. Participation was voluntary. Students who agreed to participate consented for data to be used for research.

Data sources/ measurement/Variables

Three sessions out of each academic year's 8 sessions were selected for the pilot study—first session, mid-session, and last session. Afer each session, the students rated each and every group mate using online evaluation forms hosted on a learning platform (Entrada).

The evaluation form provided clear criteria guidance and a level of standardisation of appropriate professional behaviour and attributes required of students. The form contained 5 questions on a 9-point Likert scale. The five questions covered these areas: Q1) integrity & honesty, Q2a) responsibility and participation (has good attendance; is punctual; participates appropriately; is a good team player) Q2b) responsibility & participation (accountable and committed to the successful completion of tasks assigned to you), Q3a) respect & sensitivity, Q4) compassion & empathy. The 9-point Likert scale used was also broadly grouped into the following: Scores 1–3 as "Below Expectations", 4–6 as "Meet Expectations", and 7–9 as "Above Expectations".

Study size

No sample side calculation was performed as this was a pilot study involving a small group of students. Also, only data from voluntarily participating students were analysed and reported.

Statistical methods:

Descriptive and inter/intra-rater reliability analysis were done using SPSS. Gwet's Agreement Coefficient (AC) was calculated and then categorized based on the Benchmark Scale - no agreement (<0.00), poor agreement (<0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.6), good agreement (0.61–0.8), very good agreement (0.81–0.99), perfect agreement (1). These were done first for each student group (by question and by session). The sum of the groups' agreement coefficients and categories were then calculated for each question, for Phase I and II separately and combined.

3 Results

Participants

Sixty six and eighty one Phase I and II medical students were invited to participate. Fifty-two (78.79% response rate) and 54 (66.67%) Phase I and II students volunteered to participate in the pilot study, of which, 21 (40.38%) and 24 (44.44%) are males for Phase I and II respectively. Their age range from 19 to 22 years old.

Main results

Data was collected over 3 sessions. Forty-seven students completed the forms for all 3 sessions.

As shown in Table 1, the average question scores among Phase I were generally higher than the Phase II students.

When looking at the student groups, in Phase I—3 groups and year II—2 groups respectively rated 9 for all their peers for some questions in few of the sessions. Within the same cohort, we looked at changes in the questions over time, comparing the start (first session) to the end (last session). For Phase I, Q1 had the biggest change over time. A similar trend was observed for Phase 2 with all the question scores showing an increase when comparing the start (first session) and end (last session), with the mid session recording the highest scores. However, the difference in score is relatively bigger (e.g. +1.2 for Q4 from 7.51 at the start to 7.63 at the last session).

In addition, we also looked at how the average question score changed over the 3 sessions at the individual group level within each cohort. We noticed several broad key patterns in how the average question scores changed over the sessions:

Table 1 Average score (based on a 9-point Likert Scale) for each Phase of study by question and session

Phase I	Q1	Q2a	Q2b	Q3a	Q4
First session	8.11	8.16	8.16	8.17	8.13
Mid session	8.21	8.06	8.26	8.27	8.15
Last session	8.19	8.06	8.18	8.23	8.20
Phase II	*Q1*	*Q2a*	*Q2b*	*Q3a*	*Q4*
First session	7.71	7.62	7.59	7.55	7.51
Mid session	7.85	7.88	7.87	7.90	7.89
Last session	7.78	7.71	7.69	7.65	7.63
Phase I and II combined	*Q1*	*Q2a*	*Q2b*	*Q3a*	*Q4*
First session	7.91	7.89	7.88	7.86	7.82
Mid session	8.03	7.97	8.07	8.09	8.02
Last session	7.98	7.88	7.94	7.94	7.92

- Declined from first session to the last session for most questions.
- Declined from the first session to the mid-session, then increased in the last session for most questions.
- Increased in score from first session to the mid-session, then stayed similar or increased at the last session for most questions.
- Inconsistent pattern for all the questions

With regards to reliability between students in the same group rating the same peer, overall (both Phase I and II), most student groups had Perfect Agreement (meaning higher reliability) at the first session, and all student groups had perfect agreements at the mid and last session (see Table 2). When separate analysis of Phase I and II were done, a different pattern was observed. Perfect agreement was observed for all questions in the first session and mid-session for Phase I, and for some questions in the last session. For Phase II, variable levels of agreement were seen in the first session, while perfect agreement was observed for all questions in the mid-session and final session. Phase I students were more consistent and stable in rating their peers as compared to Phase II. Questions that did not have strong relaibility ("poor agreement" or "no agreement") were Q1 and Q3a (for Phase I) as well as Q4 (for Phase II).

Table 2 Level of agreement for each question across 3 sessions for Phase I, Phase II and overall

Overall	Q1	Q2a	Q2b	Q3a	Q4
First session	Poor, very good, perfect agreement	Moderate, perfect agreement	Very good, perfect agreement	Perfect agreement	Perfect agreement
Mid-session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement
Last session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement
Phase I	*Q1*	*Q2a*	*Q2b*	*Q3a*	*Q4*
First session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement
Mid-session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement
Last session	No and perfect agreement	Perfect agreement	Perfect agreement	No & perfect agreement	Perfect agreement
Phase II	*Q1*	*Q2a*	*Q2b*	*Q3a*	*Q4*
First session	Moderate and very good agreement	Moderate agreement	Fair and very good agreement	Fair agreement	No agreement and poor agreement
Mid-session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement
Last session	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement	Perfect agreement

4 Discussions

Across many disciplines, peer assessment has been found to be part of students' learning as a formative tool, as it assists in the progressive development of skills and contents [19]. The successful introduction of peer assessment is dependent on several factors, including the type of method adopted, receiving peer assessment information and how it is used, cultural influence, and issues surrounding the anonymity and confidentiality of the feedback [12, 21]. It is crucial to investigate whether the western approach to promote professionalism will work in a non-western context, especially taking into consideration the context and cultural diversity into before implementation. This was a pilot study to explore whether repeated peer assessment by preclinical students help in improving one's professionalism, and the method's reliability as an evaluation tool.

From the average question scores observed across the sessions, it seemed that the students' level of professsionalism had improved over time as a higher score was observed at the end of the CLC sessions. This may be because some of the professionalism domains evaluated in the questionnaire (e.g. empathy, integrity, sensitivity) had been taught to students during the same period as part of the other programmes within the curriculum. Secondly, this could be due to the effect of repeated evaluations where students become more conscious of being evaluated and hence change their behaviour.

Several groups were observed to have rated their peers a perfect score of 9 in all questions, especially towards the last session. This might be due to learner improvement as discussed above, but it could also be due to peer pressure. It has been noted that peer assessments tend to have the issues of over- marking, under-marking, or friendship-marking [20]. For instance, a student may be a close friend of the peer whom he or she was evaluating and will give a high rating to help their friend have better scores. Also, as the students in the same group attend other classes and work together closely outside of CLC, they might be reluctant to penalize their fellow groupmates as it might be awkward after that to work together [20].

We noticed several broad key patterns on how the average question scores changed over the sessions. A published review suggested that while the impact of an education approach is likely due to the approach itself and some methodological limitations, it might also be influenced by individual differences (student/teacher) and contextual variabilities [22]. In recent years, our school has expanded the diversity of its medical students, for instance through enrolling applicants with various backgrounds, accessing applicants based on other traits rather than focusing entirely on academic results. These could have resulted in students with more diverse learning approaches. Furthermore, in the study carried out by Curran et al., the authors discovered that their students expressed concerns regarding the functional lack of anonymity due to the small group size. Consequently, negative feedback and rating were avoided. As our students' group size was small, this may also have contributed to the variability in the rating [10]. As for the tutors, the sessions were usually taught by different educators, and this may have affected the peer discussions and hence peer behaviour.

Moving on to the level of agreement, overall, less agreement were seen in the first session, and more perfect agreement was observed in the mid-session and last session. This was not unexpected, since at the first session, students were not familiar with the scale, and had little experience in rating their peers. We noticed that one or two questions (Q1 and Q4) were more problematic. These questions evaluated traits (i.e., integrity, empathy) which tend to be multi-faceted and highly subjective, and hence could have contributed to the variability. This outcome contradicts with the study carried out by Nofziger et al. whereby they found that peers can provide reliable and stable ratings of both work habits (e.g., preparation, problem solving initiative) and interpersonal attributes (e.g., truthfulness, respect, integrity, empathy) [14].

We observed that Phase II students had perfect agreement for all questions in the last session and this was not observed for Phase I. In addition to the factors that can influence the realiability of peer assessment (including the number of relevant performances observed, the number of peers involved and the number of aspects of competence being evaluated), student maturity may affect the rating. It is possible that Phase II students being more mature, having had more professionalism training, thus have better ethos of the processes and know what was appropriate and inappropriate [4].

This pilot study had provided us with insights on areas to improve. Firstly, we need to improve and standardize the evaluation process. Briefings will be conducted to provide more information on the goals, rating scale, items, evaluation rubric, and what areas to look out for each traits. Secondly, we need to provide sufficient training on how to provide feedback so that we can improve the accuracy and effectiveness of the feedback [23]. Thirdly, we will need to optmise the environment by ensuring a transparent evaluation culture whereby students are well aware of all components of the process [23]. Lastly, we would emphasise to students that it is okay to give a lower rating if necessary, as this is part of the formative learning process and will help their peers to identify areas for improvement.

Limitations

There are several limitations with our study. First, the sample size was small as only a portion of each cohort was chosen for the study, and not all the students from the selected groups participated. Next, the students may not have completed the form immediately after the lesson, and might have done it several days later. This could result in the issue of recall bias. Thirdly, we only did 3 sessions, thus limiting the amount of data collected. However, considering that this is a pilot study, the issues above were not unexpected.

Generalizability

While this is only a pilot study, the results have been encouraging. The school is now considering incorporating peer assessment as a standard practice across the years of study. Other schools especially those from Asia could also consider implement peer assessment in their small group teaching.

Suggestions

Future studies could focus on comparing the students' peer assessment with that of the tutors' student assessment to see if there are concordance or discordance in the assessments.

5 Conclusion

In conclusion, our pilot study has suggested that longitudinal peer assessment has the potential to be a reliable tool for assessment over repeated measures and it can help to improve the professionalism of the students over time.

Acknowledgements We would like to express our gratitude to the tutors and students involved in the study as well as Ms Wong Hung Chew from the NUS Medicine Biostatistics Unit who helped to conduct the statistical analysis for the study.

Supplementary Materials Nil.

Funding No funding.

Conflict of Interest No potential conflicts of interest relevant to this article were reported.

References

1. Hafferty F, Papadakis M, Sullivan W, Wynia MK (2012) The American board of medical specialties ethics and professionalism committee definition of professionalism. 2012 Chicago, Ill American Board of Medical Specialties
2. Mueller PS (2015) Teaching and assessing professionalism in medical learners and practicing physicians. Rambam Maimonides Med J 6(2). https://doi.org/10.5041/RMMJ.10195
3. Dannefer EF, Henson LC, Bierer SB, Grady-Weliky TA, Meldrum S, Nofziger AC, Barclay C, Epstein RM (2005) Peer assessment of professional competence. Med Educ 39:713–722. https://doi.org/10.1111/j.1365-2929.2005.02193.x
4. Finn GM, Garner J (2011) Twelve tips for implementing a successful peer assessment. Med Teach 33:443–446. https://doi.org/10.3109/0142159X.2010.546909
5. Spandorfer J, Puklus T, Rose V, Vahedi M, Collins L, Giordano C, Braster C (2014) Peer assessment among first year students in anatomy. Anat Sci Educ 7:144–152. https://doi.org/10.1002/ase.1394
6. Speyer R, Pilz W, Van Der Kruis J, Brunings JW (2011) Reliability and validity of student peer assessment in medical education: a systematic review. Med Teach 33:e572–e585. https://doi.org/10.3109/0142159X.2011.610835
7. Lerchenfeldt S, Mi M, Eng M (2019) The utilization of peer feedback during collaborative learning in undergraduate medical education: A systematic review. BMC Med Edu 19(1):1–10. https://doi.org/10.1186/s12909-019-1755-z
8. Goldie J (2013) Assessment of professionalism: a consolidation of current thinking. Med Teach 35(2):e952–e956. https://doi.org/10.3109/0142159X.2012.714888
9. Chen JY (2012) Why peer evaluation by students should be part of the medical school learning environment. Med Teach 34(8):603–606. https://doi.org/10.3109/0142159X.2012.689031

10. Curran VR, Fairbridge NA, Deacon D (2020) Peer assessment of professionalism in undergraduate medical education. BMC Med Edu 20(1):1–8. https://doi.org/10.1186/s12909-020-02412-x
11. Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM (2006) Temporal and group-related trends in peer assessment amongst medical students. Med Educ 40:840–847. https://doi.org/10.1111/j.1365-2929.2006.02540.x
12. Shue CK, Arnold L, Stern DT (2005) Maximizing participation in peer assessment of professionalism: the students speak. Acad Med 80(10):S1-5
13. Sullivan ME, Hitchcock MA, Dunnington GL (1999) Peer and self assessment during problem-based tutorials. Am J Surg 177:266–269. https://doi.org/10.1016/S0002-9610(99)00006-9
14. Nofziger AC, Naumburg EM, Davis BJ, Mooney CJ, Epstein RM (2010) Impact of peer assessment on the professional development of medical students: a qualitative study. Acad Med 85:140–147. https://doi.org/10.1097/ACM.0b013e3181c47a5b
15. Norcini JJ (2003) Peer assessment of competence. Med Educ 37:539–543. https://doi.org/10.1046/j.1365-2923.2003.01536.x
16. Thanh PT, Gillies R (2010) Designing a culturally appropriate format of formative peer assessment for Asian students: The case of Vietnamese students. Int J Educ Reform. 19(2):72–85. https://doi.org/10.1177/105678791001900201
17. Irwin H (1996) Communicating with Asia: Understanding people and customs. Allen & Unwin, St. Leon-ard's, New South Wales, Australia
18. Samarasekera DD, Lieske B, Aw D, Lee SS, Lim YH, Ang CY, Yeo SP, Koh DR (2021) A new model of teaching and learning approach—collaborative learning cases activities. Asia Pac Schol 6(2):98. https://doi.org/10.29060/TAPS.2021-6-2/MA1602
19. Planas Lladó A, Soley LF, Fraguell Sansbelló RM, Pujolras GA, Planella JP, Roura-Pascual N, Suñol Martínez JJ, Moreno LM (2014) Student perceptions of peer assessment: an interdisciplinary study. Assess Eval High Educ 39(5):592–610. https://doi.org/10.1080/02602938.2013.860077
20. Carvalho A (2013) Students' perceptions of fairness in peer assessment: evidence from a problem-based learning course. Teach High Edu 18(5):491–505. https://doi.org/10.1080/13562517.2012.753051
21. Arnold L, Shue CK, Kritt B, Ginsburg S, Stern DT (2005) Medical students' views on peer assessment of professionalism. J Gen Intern Med 20:819–824. https://doi.org/10.1111/j.1525-1497.2005.0162.x
22. Onyura B, Baker L, Cameron B, Friesen F, Leslie K (2016) Evidence for curricular and instructional design approaches in undergraduate medical education: an umbrella review. Med Teach 38(2):150–161. https://doi.org/10.3109/0142159X.2015.1009019
23. Lerchenfeldt S, Taylor TA (2020) Best practices in peer assessment: training tomorrow's physicians to obtain and provide quality feedback. Adv Med Educ Pract 11:571. https://doi.org/10.2147/AMEP.S250761