Evaluating Inter-Rater Reliability: Transitioning to a Single Rater for Marking Modified Essay Questions in Undergraduate Medical Education

Shahid Hassan[1], Malanashita Ganeson[2], Ismail Abdul Sattar Burud[3]

[1] School of Medicine, American University of Barbados, Bridgetown, Barbados
[2] Department of Family Medicine, Kualalumpur, Malaysia
[3] Department of Surgery, School of Medicine, International Medical University, Kuala Lumpur, Malaysia

Received: 04 Jan. 2024; Accepted: 23 Feb. 2024

Abstract- Modified Essay Questions (MEQs) are often included in high-stakes examinations to assess higher-order cognitive skills. If the marking guides for MEQs are inadequate, this can lead to inconsistencies in marking. To ensure the reliability of MEQs as a subjective assessment tool, candidates' responses are typically evaluated by two or more assessors. Previous studies have examined the impact of marker variance. Current study explores the possibility of assigning a single assessor to mark the students' performances in MEQ based on statistically drawn evidence in the clinical phase of the MBBS program at a private medical school in Malaysia. A robust evaluation method was employed to determine whether to continue with two raters or shift to a single-rater scheme for MEQs, using the Discrepancy-Agreement Grading (DAG) System for evaluation. A low standard deviation was observed across all 11 pairs of scores, with insignificant t-statistics (P>0.05) in 2 pairs (18.18%) and significant t-statistics (P<0.05) in 9 pairs (81.81%). The Intraclass Correlation Coefficient (ICC) results were excellent, ranging from .815 to .997, all with P<0.001. Regarding practical effect size (Cohen's d), 1 pair (9.09%) was categorized as having a strong effect size (>0.8), 7 pairs (63.63%) as having a moderate effect size (0.5-<0.8), and 3 pairs (27.27%) as having a weak effect size (0.2-<0.5). The data analysis suggests that it is feasible to consider marking MEQ items by a single assessor without negatively impacting the reliability of the MEQ as an assessment tool.

Acta Med Iran 2024;62(March-April):88-95.

Keywords: Essay question; Decision making; Observers variation; Interobserver reliability; Scoring system

Introduction

Modified Essay Questions (MEQs) and Restricted Response Essay (RRE) questions are widely used assessment tools in both undergraduate and postgraduate medical education. MEQs, in particular, are employed to evaluate students' integrated and holistic learning experiences within the curriculum. This format allows for the creation of interdisciplinary assessment questions, enhancing the overall educational evaluation process. However, due to the subjective nature of MEQs, candidates' responses are typically assessed by two or more evaluators, with or without the aid of model answers for each clinical attribute (or section) of the MEQ (Figure 1). Often, the final score for each section is determined through discussion among the evaluators to reach a consensus, thereby minimizing discrepancies and disagreements. While these measures are intended to reduce subjective bias and enhance the reliability of assessments, the process is often time-consuming, involving multiple evaluators and several steps in the marking process. The subjectivity inherent in faculty judgments when assessing essay questions can introduce bias, leading to the failure to document significant deficits and limiting the discrimination between different levels of student performance (1). Potential sources of judgment errors include the individual rater, the rating scales, the items being rated, and the objects of rating (2).

Corresponding Author: Sh. Hassan
School of Medicine, American University of Barbados, Bridgetown, Barbados
Tel: +98 60129539853, E-mail address: gorshahi@gmail.com

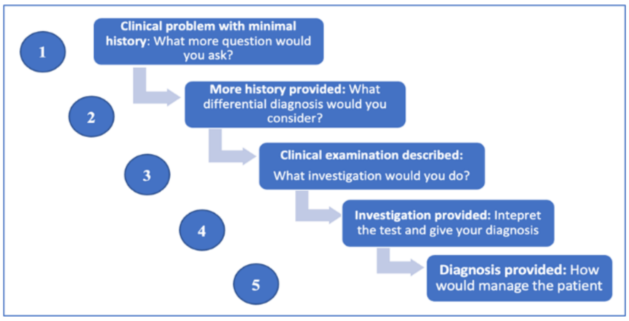

1
Clinical problem with minimal history: What more question would you ask?
2
More history provided: What differential diagnosis would you consider?
3
Clinical examination described: What investigation would you do?
4
Investigation provided: Intepret the test and give your diagnosis
5
Diagnosis provided: How would manage the patient

Figure 1. Algorithm showing sequence of Modified Essay Question (MEQ) administered and withdrawn one after another

These challenges can compromise the raters' ability to maintain consistent standards. To mitigate rater bias, multiple assessors are often involved in marking essay questions. Previous studies have investigated the effects of marker variance, finding that ambiguous wording and incomplete marking schemes account for significant proportions of item-writing and marking scheme flaws (3). Despite the use of multiple markers, reliability gains are minimal; for example, using two markers only slightly increases Cronbach's alpha coefficient. It has been shown to be more cost-effective to single-mark two questions rather than double-mark one (4). Moreover, studies indicate that increasing the number of questions on a MEQ exam can significantly enhance reliability, more so than increasing the number of assessors (5). Generalizability theory is a technique that extrapolates existing data to predict the reliability which would be achieved by changing the structure of the examination paper. The application of generalizability theory indicated that by increasing the number of questions, the reliability of the paper could be greatly enhanced rather than increasing the number of assessors (6). In medical education, the reliability of assessments is crucial for accurately evaluating learner performance and upholding educational standards. Rater judgments, a critical component of many assessment methods, present significant challenges due to their inherent variability and potential for error. Maintaining the reliability of these judgments is essential for preserving the validity of assessments and, consequently, the quality of medical education. Research highlights the complexities associated with rater judgments, noting that variability among raters can undermine the consistency and fairness of evaluations. This underscores the need for robust strategies to improve rater reliability and ensure accurate assessment outcomes.

Albanese (2000) also outlines the complexities associated with rater judgments in medical education, noting that variability among raters can undermine the consistency and fairness of evaluations (1). Downing (2005) further underscores this issue by highlighting the threats to validity posed by rater error in clinical teaching assessments (2). These concerns emphasize the need for robust strategies to improve rater reliability and ensure accurate assessment outcomes. Palmer *et al*., (2010) explore the MEQ's effectiveness and its evolution in exit examinations, revealing both its strengths and limitations (4). The consistent application of this format is crucial, as variability in rater judgments can significantly impact the assessment's reliability. To accurately measure and interpret rater reliability, statistical methods are essential. Landis and Koch (1977) and Shrout and Fleiss (1979) provide fundamental insights into intraclass correlations and other statistical techniques for assessing rater agreement (7,8). Additionally, Streiner and Norman (2008) offer a practical guide to health measurement scales, including strategies for improving the reliability and validity of assessments to reduce the subjective bias (9). Recent advancements in grading systems, such as the Discrepancy-Agreement Grade (DAG) proposed by Yusoff and Rahim (2012), aimed to address discrepancies to improve feedback on rater judgments (10). This novel system provides a more nuanced approach to evaluating rater performance and ensuring accurate feedback. Finally, understanding the practical significance of research findings is crucial. Lakens (2013) provides a comprehensive guide to calculating and reporting effect sizes for t-tests and ANOVAs, which is vital for interpreting the impact of interventions aimed at improving rater reliability though out of the scope of current study (11).

The effectiveness of the MEQ format in exit examinations has been explored in various studies, revealing both strengths and limitations. Consistent application of this format is vital, as variability in rater judgments can significantly impact the reliability of assessments. To accurately measure and interpret rater reliability, statistical methods such as intraclass correlation coefficients (ICCs) and Cohen's Kappa are essential. These techniques provide valuable insights into rater agreement and the overall reliability of assessments. Recent advancements in grading systems, like the Discrepancy-Agreement Grade (DAG), have been developed to address discrepancies in rater judgments, offering a more nuanced approach to evaluating rater performance and ensuring accurate feedback.

The rationale for the current study is to address critical issues in rater reliability, particularly the subjective nature of MEQs, which are designed to evaluate higher-order cognitive skills such as clinical reasoning, decision-making, and problem-solving. Variability in marking can significantly influence student performance, and this study aims to explore whether a single rater can provide reliable and consistent scores, thereby improving resource efficiency. By examining the effectiveness of various strategies and statistical methods, this research seeks to contribute to the ongoing improvement of assessment practices in medical education, ensuring more accurate and reliable evaluations of learner performance.

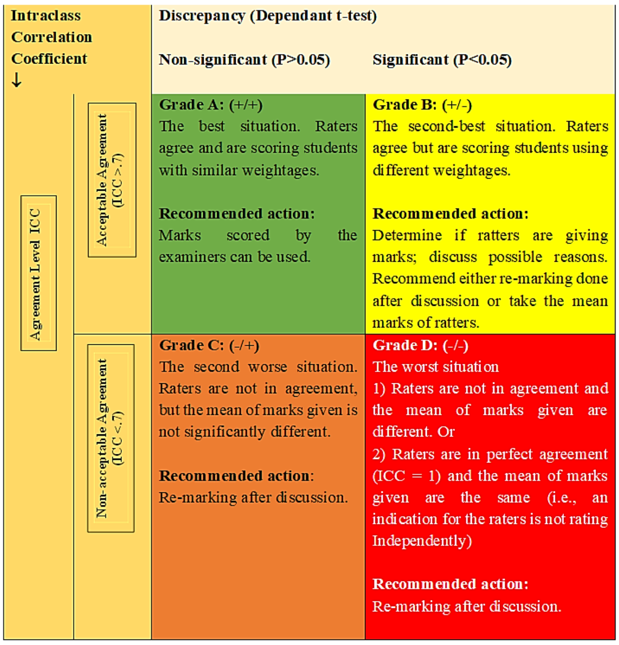
Statistical evidence is crucial in justifying the use of a single rater. Methods such as Intraclass Correlation Coefficient (ICC) for inter-rater agreement, Cohen's Kappa for categorical items, Cohen's d for effect size, t-tests, and ANOVA for significant differences, and the Discrepancy-Agreement Grading (DAG) system can be employed to quantify and analyse any discrepancies between raters. By thoroughly examining inter-rater reliability using these statistical tools, educational institutions can make informed, evidence-based decisions about whether MEQs can be marked by a single evaluator without negatively impacting student performance scores. This approach ensures that the assessment process remains fair, reliable, and efficient, thereby maintaining the integrity of the examination process.

Materials and Methods

A cross-sectional study conducted on eleven pairs of faculty assessors who rated Modified Essay Questions (MEQ) in End of Semester (EOS) 9 Examination. The items submitted by combined two disciplines to create MEQ as an integrated approach and reviewed by a discipline and central vetting committee attended by subject and some non-subject experts were followed through. Questions were vetted as usual for clarity relevancy and appropriateness of weighting and accuracy of the marking scheme using a model answer. Two assessors were assigned to mark the relevant component of MEQ. The data was collected as the ratting score of 6 MEQ items (parts) marked by 11 pairs of examiners, each marking independently. Each pair of MEQ examiners rated answer sheets of 114 medical students. The discrepancy and agreement level between two examiners for each pair were analysed and graded using, Discrepancy-Agreement Grading (DAG) system for this evaluation. A robust method of evaluation is used for a logical decision to either continue with two raters and their discussion to agree on a consensus ratting of student's performance or move to one maker scheme in MEQ.

DAG grids comprise of two statistical methods of Intraclass Correlation Coefficient (ICC) to

determine the level of agreement between the two raters set a minimal value of ICC=0.7 and dependent or paired t-Test to determine the level of significant of mean score between the two raters. The Discrepancy-agreement Grade (DAG) system is a two-way statistical method developed to measure inter-rater variability. Two statistical tests involved are the paired or dependent-t and intraclass correlation (ICC). The dependent-t test and ICC help to determine discrepancy and agreement between two raters respectively. The discrepancy in mean ratting score between the two raters was considered non-significant if *P* of the dependent-t test is more than 0.05. On the other hand, the level of agreement is considered as acceptable at a value of 0.7 using ICC, whereas in the original DAG System ICC value is set at 0.4) (10). Based on the results of the two tests ratting, judgments are classified into grades A, B, C or D (see figure 2). Grade A is considered the best condition based on DAG system where the two raters have a good agreement level and are scoring with similar weightages. Grade B is the condition where the two raters are scoring with different weightages but have a good agreement and a mean or discussion to reach a consensus mark of both raters is recommended. Grade C is the condition where the two raters have a poor agreement level but with no obvious discrepancy of mean marks given. Grade D is considered worst condition where the two raters are in poor agreement and scoring with dissimilar weightages. A remarking after discussion is recommended for grades C and D. Grade A and B are considered good level while grade C and D are considered as a poor level of rater judgments (see figure 2).



Intraclass Correlation Coefficient ↓		Discrepancy (Dependant t-test)	
		Non-significant (P>0.05)	Significant (P<0.05)
Agreement Level ICC	Acceptable Agreement (ICC >.7)	**Grade A:** (+/+) The best situation. Raters agree and are scoring students with similar weightages. **Recommended action:** Marks scored by the examiners can be used.	**Grade B:** (+/-) The second-best situation. Raters agree but are scoring students using different weightages. **Recommended action:** Determine if ratters are giving marks; discuss possible reasons. Recommend either re-marking done after discussion or take the mean marks of ratters.
	Non-acceptable Agreement (ICC <.7)	**Grade C:** (-/+) The second worse situation. Raters are not in agreement, but the mean of marks given is not significantly different. **Recommended action:** Re-marking after discussion.	**Grade D:** (-/-) The worst situation 1) Raters are not in agreement and the mean of marks given are different. Or 2) Raters are in perfect agreement (ICC = 1) and the mean of marks given are the same (i.e., an indication for the raters is not rating Independently) **Recommended action:** Re-marking after discussion.

Figure 2. The Discrepancy-Agreement Grade (DAG) Grid: showing possible results of data analysis and recommended actions: Adopted from Muhamad Saiful and Ahmad Fuad, 2012

Results

Mean and SD as well as t-statistics of dependant t-test and the F-statistics of ICC values were analysed for their significance levels of *P*=0.05 and applied to DAG system for interpretation. The discrepancy and agreement level between the two examiners for each pair were analysed and graded based on the DAG grid (see figure 2).

On eyeball rolling a minor difference in mean with low SD (see table 1) was observed for all 11 pairs scores with insignificant t-statistics (*P*=>0.05) in 2 (18.18%) and significant t-statistics (*P*=<0.05) in 9 (81.81%) out of 11 pairs (see table 2). ICC result was excellent varying between .815 to .997 and with their *P*=<0.001 in all 11 Pairs of assessors involved in the end of semester examination-9 of MEQ examination (see table 3).

The narrow confidence interval of <1 between the upper and the lower values of of Dependent t-Test Statistics and Intraclass Correlation Coefficient showed a high precision in the estimate (see table 1 and table 2)

Observing the practical effect size as Cohen’s d, only 1 (9.09%) was categorised strong (>0.8) vs. 7 (63.63%) categorised moderate (0.5-<0.8) and 3 (27.27%) categorised weak (0.2<0.5) (see table 3)

Table 1. Mean and standard deviation of paired sample t-test of two raters

No. of Pair	No. of Rater	Mean	Std. Deviation
Pair 1	Rater 1	2.355	1.412
	Rater 2	3.053	1.585
Pair 2	Rater 1	2.535	1.509
	Rater 2	2.583	1.510
Pair 3	Rater 1	4.700	1.645
	Rater 2	4.143	1.642
Pair 4	Rater 1	3.020	1.463
	Rater 2	3.700	1.527
Pair 5	Rater 1	4.890	1.400
	Rater 2	5.160	1.367
Pair 6	Rater 1	6.350	1.185
	Rater 2	6.557	.918
Pair 7	Rater 1	6.257	1.308
	Rater 2	6.322	1.281
Pair 8	Rater 1	4.191	1.968
	Rater 2	4.352	2.019
Pair 9	Rater 1	4.640	2.295
	Rater 2	4796	2.165
Pair 10	Rater 1	3.735	2.114
	Rater 2	3.848	2.127
Pair 11	Rater 1	3.678	2.044
	Rater 2	3.734	2.144

Table 2. Significance Level of Dependent t-Test Statistics

Pair/Rater	t statistics	df	Significance	95% CI Lower	95% CI Upper
Pair1/Rater 1,2	-9.308	113	<.001	-.845	-.548
Pair2/Rater 1,2	-1.115	114	.251	-.1299	.0342
Pair3/Rater 1,2	5.743	114	<.001	.3617	.742
Pair4/Rater 1,2	2.900	114	.004	.099	.527
Pair5/Rater 1,2	-.3526	114	<.001	-.421	-.118
Pair6/Rater 1,2	-4.099	114	<.001	-.3096	-.1078
Pair7/Rater 1,2	-2.133	114	.035	-.1528	-.0046
Pair8/Rater 1,2	-.2624	114	.010	-.2823	-.0394
Pair9/Rater 1,2	-.3.150	114	.002	-.2479	-.0565
Pair10/Rater 1,2	-3.233	114	.002	.-1823	-.0438
Pair11/Rater 1,2	-.971	114	.333	-.1717	.0587

Table 3. Intraclass correlation coefficient and the level of significance

Pair/Rater	ICC	95% CI Lower	95% CI Upper	F statistics	Significance
Pair1/Rater 1,2	.977	.937	.969	44.097	<.001
Pair2/Rater 1,2	.992	.974	.989	128.784	<.001
Pair3/Rater 1,2	.985	.977	.990	73.204	<.001
Pair4/Rater 1,2	.971	.957	.980	35.806	<.001
Pair5/Rater 1,2	.983	.975	.988	61.357	<.001
Pair6/Rater 1,2	.924	.975	.952	14.914	<.001
Pair7/Rater 1,2	.895	.837	.931	10.396	<.001
Pair8/Rater 1,2	.815	.726	.874	5.677	<.001
Pair9/Rater 1,2	.868	.733	.926	9.417	<.001
Pair10/Rater 1,2	.978	.968	.985	45.234	<.001
Pair11/Rater 1,2	.874	.495	.948	13.094	<.001

Discussion

The Modified Essay Question (MEQ) format is widely utilized in medical education to assess clinical reasoning and knowledge. Traditionally, the MEQ introduces a patient scenario presented through a step-by-step narrative, where medical problems are gradually revealed, simulating the uncertainties and time constraints typically encountered during a clinical assessment (see Figure 1). Discrepancy-Agreement Grading (DAG) offers pertinent feedback regarding faculty grading performance marked by two evaluators, which can serve as evidence to enhance their future rating accuracy and to make informed decisions about assigning a single rater for marking MEQs. Numerous studies support the idea that feedback on assessors rating is a powerful tool for improving individual performance (12-14) and can be used as evidence for transforming to one evaluator. The feedback provided through DAG gave faculty valuable insights into their evaluation of student performance.

In the current study, two evaluators were assigned to grade the MEQs, with each pair of examiners assessing different components of the same question. Six MEQ items were rated by eleven pairs of faculty members during the end-of-semester 9 examination. Among these pairs, 2 (18.18%) received an "A" grade, while 9 (81.82%) received a "B" grade. These results indicate that there was a significant level of agreement between the examiners, albeit with some discrepancies. In the context of DAG, a discrepancy refers to a situation where two markers are linearly related but show inconsistency in their ratings, determined by a certain degree of difference. Agreement, on the other hand, occurs when the ratings are not only linearly related but also consistent in assigning the same scores and overall grades. Discrepancies suggest differing weightings, whereas agreement implies similar weightings. These differences can impact student outcomes due to rater errors. An extreme example could be two raters who consistently score a student with a 4/10 and 5/10, respectively, yet disagree on the final outcome, with one assigning a pass and the other a failure, based on a 50% cut-off score.

Developing an MEQ paper can be relatively straightforward, but marking it is often more onerous and challenging (12). To ensure the reliability of MEQ, which is a subjective assessment, candidates' responses were marked by two assessors online using the model answers. Faculty involved in this process found it unnecessary, time-consuming, and labour-intensive, suggesting that a single assessor should mark the MEQs. However, assigning one assessor to mark each component of the MEQ requires not only sufficient evidence but also a robust and convincing method for decision-makers. A high level of agreement between the two raters indicated an excellent intraclass correlation coefficient (ICC) above .8 in all 11 pairs (see Table 3). Although there were some discrepancies in ratings among the pairs, a high level of agreement was observed between them. The consistently high intraclass correlation coefficients across all raters, along with significant differences in mean scores by most pairs, ultimately led all pairs of assessors to be placed in grades A and B on the DAG grid (see Table 4). The data analysis of examiners' rating scores on candidates' performances provides sufficient evidence to suggest that moving forward with marking MEQ items by a single assessor is feasible without compromising the reliability of the assessment, as indicated by the highly significant intraclass correlation coefficient values.

On the other hand, comparing the practical or clinical effect with the statistically significant difference between the mean rating scores of the two raters helps determine the magnitude of the effect. Cohen's d, or standardized mean difference, is one of the most common ways to measure effect size. Determining the effect size in data analysis quantifies the magnitude of a relationship or difference between groups, independent of the sample size. While a *P* can indicate whether an effect exists, it does not reveal the size of that effect. Cohen's d can be used as an effect size statistic for a paired t-test. Additionally, the narrow Confidence Interval (CI) of less than 1 between the upper and lower values of the dependent t-Test statistics and ICC indicated high precision in the estimate, suggesting that the true difference is very close to the observed value. This high precision enhances confidence in the reliability of the results, indicating that the observed difference is consistent and robust. A narrower 95% CI provides stronger evidence to support the decision to adopt a policy of moving to a single evaluator for marking MEQ items.

Including effect size in the discussion ensures a more comprehensive evaluation of the study's findings, facilitating a deeper understanding of their implications for educational practices in medical schools. In the current study, the effect size as measured by Cohen's d was mostly categorized as having a moderate to weak effect size, except in one pair, which showed a strong effect size. Cohen's (1988) proposed a Cohen's d formula (15) for the interpretation of effect size (see Table 5). Understanding the magnitude of any differences helps in assessing whether adopting a single rater is likely to have meaningful consequences on the fairness and accuracy of student assessments. Effect size aids in making informed decisions about the potential benefits or drawbacks of implementing single rater assessments, ensuring that decisions are based on the magnitude of effects rather than just their statistical significance.

Table 4. DAG grade and the effect size versus significance.

Pair/Rater	DAG Grade	Effect Size Cohen's d Paired Sample t-test	t-Test Significance (2-tailed)
Pair1/Ratter 1,2	B	.872 (Strong)	<0.001
Pair2/Rater 1,2	A	.108 (Very weak)	.251
Pair3/Rater 1,2	B	.536 (Moderate)	<0.001
Pair4/Rater 1,2	B	.270 (Moderate)	0.004
Pair5/Rater 1,2	B	.329 (Moderate)	<0.001
Pair6/Rater 1,2	B	.382 (Moderate)	<0.001
Pair7/Rater 1,2	B	.199 (Very weak)	.035
Pair8/Rater 1,2	B	.245 (Moderate)	.010
Pair9/Rater 1,2	B	.294 (Moderate)	.002
Pair10/Rater 1,2	B	.301 (Moderate)	.002
Pair11/Rater 1,2	A	.091(Very weak)	.331

Table 5. Cohen's d formula and interpretation of effect size (Cohen,1988)

Cohen's d Formula	No	Cohen's d	Interpretation
Mean1 – Mean2	1	0.2 < 0.5	Small effect
Pooled SD	2	0.5 < 0.8	Medium effect
	3	0.8 < 1.20	Large effect

A high ICC with significant differences but low to moderate effect sizes by the majority of pairs ultimately resulted in grades A and B in the DAG grid for all pairs. The data analysis of examiners' ratings on candidates' performances provides adequate evidence to suggest that transitioning to marking MEQs by a single assessor is unlikely to significantly impact the passing rate. However, it is recommended that if a single marker is assigned, mandatory faculty briefing and training in calibration methods should be conducted to minimize differences between raters. Additionally, a standardized rubric with model answers should be used to mark the MEQs. The findings also indicate that there will be no impact on the passing rate when assessments are conducted by a single rater using the MEQ assessment tool.

The results show that although there were statistically significant differences in mean scores between pairs of raters, the discrepancies were minor to moderate. Importantly, the high levels of agreement, as reflected by intraclass correlation coefficients (ICC) above .8 across all pairs, suggest strong consistency between raters, even when minor discrepancies are present. The application of the Discrepancy-Agreement Grading (DAG) system in this study provided a nuanced analysis of rater agreement, offering more depth than traditional inter-rater reliability metrics. This approach allowed for a rigorous assessment of the potential impact of transitioning to a single-rater system on grading accuracy and fairness. The narrow confidence intervals observed in the dependent t-test statistics and ICC indicate high precision in the estimates, further enhancing the reliability of the results. This precision suggests that the true differences in rater judgments are likely close to the observed values, supporting the robustness of our findings.

When comparing our results with existing literature, previous studies have shown that while two-rater systems are traditionally employed to enhance reliability in subjective assessments like MEQs, the burden on faculty workload is significant, particularly in the context of increased responsibilities due to remote curriculum delivery during the COVID-19 pandemic. Our findings align with the argument that a single-rater system, when properly calibrated and supported by standardized rubrics and model answers, can maintain assessment quality while reducing faculty workload. Studies by van der Vleuten *et al.*, (1996) and Hutchinson *et al.*, (2013) have similarly reported that the use of standardized rubrics and thorough rater training can mitigate the risk of bias and inconsistency in single-rater assessments, ensuring fairness and reliability (16-17).

Moreover, the calculated effect sizes, primarily categorized as moderate to weak, suggest that the differences between raters, while statistically significant, are not practically large enough to warrant concern. This finding is consistent with Norman and Schmidt (1992), who also observed that effect sizes in educational assessments often reveal that statistically significant differences do not always translate to meaningful discrepancies in educational outcomes (18). In conclusion, our study provides evidence that transitioning to a single-rater system for MEQ assessment is feasible and unlikely to negatively impact student outcomes, particularly with proper rater training and standardized assessment tools. Future research should continue to explore the long-term implications of this approach, particularly in diverse educational settings, to ensure the generalizability of these findings.

References

1. Albanese MA. Challenges in using rater judgments in medical education. Acad Med 2000;75:975-80.
2. Downing SM. Threats to the validity of clinical teaching assessments: What they are and what to do about them. Med Educ 2005;39:249-255.
3. Mulholland H, McAleer S Report to the Examination Board of the Royal College of General Practitioners. Dundee. Centre for Medical Education. Unpublished; 1988
4. Palmer EJ, Duggan P, Devitt PG, Russell R. The modified essay question: Its exit from the exit examination? Med Teach 2010;32:e300-7.
5. Lockie C, McAleer S, Mulholland H, Neighbour R, Tombleson P. Modified essay question (MEQ) paper: Perestroika. Occas pap 1990:46;18-22.
6. Brennan RL. Elements of Generalizability Theory. Iowa, ACT Publications; 1983.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
8. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
9. Streiner DL, Norman GR. Health Measurement Scales: A practical guide to their development and use. 4th ed. Oxford, England: Oxford University Press; 2008.
10. Yusoff MSB, Rahim AFA. "Discrepancy-agreement grading method: An alternative grading method to assess the assessment discrepancy in multiple choice questions. Educ Med J 2012;4:e22-33.
11. Lakens D. Calculating and reporting effect sizes to

facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863.

12. Hattie J, Timperley H. The power of feedback. Rev Educ Res 2007;77:81-112.
13. Kluger AN, DeNisi A. Feedback interventions: Toward the understanding of a double-edged sword. Curr Dir Psychol Sci 1998;7:67-72.
14. Norcini J. The power of feedback. Med Educ 2010;44:16-7.
15. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. United States: Lawrence Erlbaum Associates;1988.
16. van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ 1991;25:110-118.
17. Hutchinson L, Marks T, Pittilo M. The effects of standardized instructions on rater reliability in subjective assessments. Med Teach 2013;35:391-5.
18. Norman GR, Schmidt HG. The psychological basis of problem-based learning: A review of the evidence. Acad Med 1992;67:557-65.