Education in Medicine Journal

ORIGINAL ARTICLE

Title: Comparison of Academic, Administrative and Community Rater Scores at a Multiple Mini-Interview Using Generalisabilty Theory

Authors: Chew-Fei Sow, Carlos Collares, Allan Pau, Cees Van der Vleuten

Submitted Date: 07-10-2021

Accepted Date: 17-01-2023

Please cite this article as: Sow C-F, Collares C, Pau A, Van der Vleuten V. Comparison of academic, administrative and community rater scores at a multiple mini-interview using generalisabilty theory. Education in Medicine Journal. (Early view)

ORIGINAL ARTICLE

ARTICLE INFO

Submitted: 07-10-2021
Accepted: 17-01-2023

Comparison of Academic, Administrative and Community Rater Scores at a Multiple Mini-Interview Using Generalisabilty Theory

Chew-Fei Sow[1], Carlos Collares[2], Allan Pau[3], Cees Van der Vleuten[2]

[1]*Clinical Skills and Simulation Centre, International Medical University, Malaysia*
[2]*School of Health Professions Education, University of Maastricht, Netherlands*
[3]*Division of Community and Child Oral Health, School of Dentistry, International Medical University, Malaysia*

To cite this article: Sow C-F, Collares C, Pau A, Van der Vleuten V. Comparison of academic, administrative and community rater scores at a multiple mini-interview using generalisabilty theory. Education in Medicine Journal. (Early view)

ABSTRACT

Multiple-Mini-Interviews (MMIs) are sampling approaches that use multiple short stations to select prospective students for professional programmes. Each station uses different interview scenarios and raters to effectively assess candidates' noncognitive skills. This study compared the performances of three sets of raters; academic, administrative staff, and community members, in an MMI for student selection using performance comparisons and Generalisability Theory to estimate the different sources of variance and generalisability (reliability) coefficients. The study aims to analyse the differences in performance scores from these raters and their psychometric projections on reliability with different samples of raters and stations. Eleven candidates participated in the 10-station MMI, each with an eight-minute duration, two minutes of preparation, and an academic assessment using a marking rubric. The entire interview was video recorded. The administrative staff and community members watched the videos independently and graded all candidates' performances using the same marking rubric. Generalisability and decision studies were used to analyse the collected data. Community members were the strictest, while academics were the most lenient. There were statistically significant differences between rater categories in 6 stations. The Generalisability coefficient of 0.85 of one-rater results from the Decision study suggested good reliability of the 10-station MMI. The Decision study found that generalisability coefficients improved more with an increasing number of raters rather than number of stations. Four stations contributed to unreliability in each rater category and a combination of the rater categories. Information on number of stations, number of raters, and type of rater combination required to achieve good reliability enabled informed decisions on the process and implementation of the MMI. The station simulation that influenced unreliability helped us improve station writing and identify focus areas for training and development.

Keywords: *Multiple-mini-interview, Generalisability-Theory, Compare assessors*

CORRESPONDING

Chew-Fei Sow, Clinical Skills and Simulation Centre, International Medical University, 126 Jalan Jalil Perkasa, Bukit Jalil, 57000 Kuala Lumpur, Malaysia
Email: chewfei_sow@imu.edu.my

INTRODUCTION

The community's involvement in providing perspectives on issues related to health professions education is not uncommon [1,2,3,4]. However, involving community members in selecting students for health professions training is less common. Increasingly, such student selection is expected to be socially responsible, i.e., selecting students representative of the communities they will serve on completion of training [5]. To this end, the involvement of community members in student selection has been researched [6,7,8]. These have tended to assess the agreement or correlation between academic and community members' scores of candidates' performances at an interview.

Interviewing medical school candidates is now a common practice, and this procedure has been reported to enhance student selection in terms of acceptability and reliability [9,10]. However, evidence exists for the task variability of any interview performance, also called context specificity [11]. Therefore, like the traditional interview, a single task is not reliable, even with perfect inter-rater reliability. Reliability is improved by sampling multiple tasks, such as Multiple Mini Interviews (MMI). MMI are multiple sampling approach to personal interviews [12, 13]. The MMI consists of multiple focused encounters intended to assess many cognitive and noncognitive skills which are inadequately assessed by the personal interview. MMI has the benefit of reducing the impact of chance and interviewer situational biases. The MMI is comprised of a considerable number of short stations, each with its own examiner/s. It offers the advantage of being flexible in terms of station development. Rater variations average out when different raters are used in different tasks. Therefore, the MMI does not need to be completely objective. A review comparing several studies between subjective and objectified measurement methods indicates that objectified methods do not intrinsically provide more reliable scores [14]. Literature has shown that subjective ratings can be reliable and valid estimates of an individual's abilities [12]. The MMI is intended to assess noncognitive skills and does not require any specific medical knowledge but instead evaluates the candidate's ability to work through a process and express their ideas logically. It usually involves discussing socio-ethical dilemmas with an interviewer, role-playing with an actor or simulated patient, or completing a given task. Each station is designed to assess a specific outcome on noncognitive skills, such as critical thinking and communication skills [15].

The use of community members in the MMI has been reported [6,7,8]. Despite the fact that the question of whether academics and community members assign performance evaluations is consistent, "no effort has been made to determine the rating tendencies of interviewers with different characteristics" [6]. Eva et al. [6] found that community members scored interviewees marginally higher than academics. Bateman et al. [8] compared the numerical interview scores awarded by the academic and community member representatives and found no statistical difference in the scores of both raters. Although both studies suggest that the involvement of community members from the local community was feasible and considered important, very little is known about the performance of the community member in a student selection interview from a psychometric perspective. Therefore, in this study, we aimed to compare the performances of three different categories of raters, i.e. academic, community member, and administrative staff interviewers, in an MMI for student selection in a medical programme using performance comparisons and Generalisability Theory (G-Theory) to estimate the different sources of variance and generalisability coefficients (reliability coefficients). We address the following three research questions:

1. Are there differences in performance scores from different rater categories?
2. What is the reliability of the performance scores on MMI according to Generalisability Theory?
3. What is the impact of different numbers of raters and stations on reliability?

International Medical University Admissions Process

The current admissions process at International Medical University for the five-year medical programme is divided into two phases. There are two intakes per year, with 200 students for each intake. An Admissions team is in charge of overseeing the entire application process. The first phase takes into consideration the relevant combination of subjects, qualifications currently/previously performed by candidates, personal statements, and references. Candidates that meet the minimum requirements will advance to the next round. The second phase is a semi-structured interview process in which two medical school academic members interview candidates for about 30 minutes. The interviewers ask questions about a range of agreed-upon topics, such as assessing personal qualities/resilience/empathy based on scenarios, medical school preparation and motivation, interpersonal and communication skills, and English language proficiency. Each interviewer assesses the candidate in each domain based on the interview, first individually and then jointly. The interviewers are requested to make qualitative comments on the candidates' interview performances. The admissions team analyses the scores and comments. The interview procedure determines the decision on accepting or rejecting the application. The current admission procedure using a semi-structured interview is time-consuming in view of the large number of candidates per year due to the two intakes annually. As a result, the institution looked into the MMI as a potential alternative student selection method, as MMI requires fewer raters' hours [12].

METHODS

Sample Selection

This study was conducted at the International Medical University as a pilot project. The exclusion criteria were for those candidates who applied for the medical programme but did not meet the pre-admission academic performance criteria. Candidates who accepted the personal interview were approached to participate in this pilot MMI study. Invitations were sent out along with information on the MMI and offered the opportunity to clarify any doubts. Potential participants were assured that the decision on the personal interview outcome for admission into the medical programme was independent of their performance in the MMI. The MMI was arranged on the same day as the personal interview for those who consented to participate. Written consent was obtained.

Data Collection

Eleven candidates participated in the 10-station MMI, each station with an eight-minute duration, with one interviewer and two minutes of preparation. The stations covered attributes including coping with stress, teamwork, altruism, honesty and integrity, and empathy. One station required the candidate to complete a task with a helper; two required the candidate to role-play with an actor; the rest consisted of one-to-one interviews with an academic (academic). All academics with prior experience in the student selection process were invited to take part in this MMI pilot study. The first ten respondents were selected. The academics were junior and senior academic members who were either

scientifically or medically qualified. All academics attended training before the start of the MMI. They were briefed on the MMI process and ensured they understood the context of the attributes to be assessed at their stations. Each station was provided with a description of the station and a marking rubric. The academics were asked to evaluate and score the candidates' performance using a rating scale. Each station's score was racticed ed to 10, resulting in a possible total aggregate maximum score of 100. The academics were also asked to provide an overall impression as well as indicate any concerns about inappropriate behaviours like being overly aggressive, timid, rude, immature, etc. The entire interview was video recorded.

Subsequently, the videos were watched by two categories of raters, i.e. non-academic university administrative staff (administrative) and racticed ed and real patients (community). The administrative category was represented by Marketing, Student Services, Academic Services, and Examinations Support Services staff. Although this category is technically community members, they have some experience within the academic world and have a better knowledge of academic expectations and student issues from the perspective of a non-content expert. The community category was represented by community members recruited as simulated patients and patients from an Outpatient Clinic. They came from various backgrounds, such as retired lawyers, retired accountants, school teachers, and non-professional jobs. Potential administrative and community raters were invited to participate, and the first ten respondents for each category were selected. The selected administrative and community raters were given training before they assessed the candidate's performance in the videos. They were given a briefing on the overall MMI process as well as the context of the attributes and the interview rating scale for the station they were rating. (See Appendix 1: sample rating scale on station 3: Coping with stress). All three rater categories racticed calibrating their ratings by watching a pre-recorded video of a sample candidate's performance. An MMI station was allocated to each rater from the administrative and community categories. In their assigned MMI station, each rater individually watched and scored all 11 candidates' performances in their station using the same rating scale used by the academic.

Data Analysis

To address the first research question, we calculated the mean scores with 95% confidence intervals across all candidates at each station and across stations using SPSS version 22. We used a one-way ANOVA to test the significant differences between the three categories of raters according to stations. To address the second and third research questions, a Generalisability Theory analysis was performed using EduG 6.1 [16], which is an open-access software freely available for download [17], for the calculation of generalisability coefficients. A Decision study was also performed using the same software to address the third research question by estimating the impact of different numbers of stations and raters on the reliability estimates.

Generalisability Theory

Generalisability Theory (G-Theory) is used to assess the consistency or dependability of scores over randomly parallel replications of a particular measurement. The G-Theory is used to estimate the magnitude of various sources of error in observed scores and the relationships among such sources [18]. Each score set is a sample consisting of all possible observations on an object of measurement. Each characteristic of measurement is defined as a facet (e.g., stations, raters). Variability in the facets can be a potential source of measurement error. The primary advantage of G-Theory is the estimation of multiple sources of error in one reliability estimate for a given situation. It also provides valuable information for the optimisation of measurement designs. For these purposes, G-Theory consists of a two-step analysis: the Generalisability study and the Decision study. The Generalisability study estimates the sources of variance influencing the measurements (variance between candidates,

stations, and raters). In contrast, the Decision study is the estimation of reliability indices as a function of concrete sample size (number of stations, number of raters, etc.) [19].

The Generalisability study is used to assess the reliability, and the result is used for performing a Decision study to find the optimal conditions for a particular measurement design. An optimised design minimises undesirable sources of error and maximises generalisability. The generalisability coefficient, under a relative, norm-referenced perspective, is a reliability coefficient that provides an estimate of the generalisability of scores for interpretation that have relative meaning (scores have meaning only in relation to each other). In an absolute, domain-referenced perspective, the generalisability coefficient is also called the dependability coefficient, and it is used when the interpretation of scores must not depend on their relative position to other scores. Therefore, a Generalisability study may use the generalisability coefficient or dependability coefficient (for relative or absolute interpretation of scores) [20, 21]. Subsequently, a Decision study may use both generalisability or dependability coefficients, depending on the relative or absolute uses and interpretations of test scores. Figure 1 summarises the overview presentation of G-Theory. In the interpretation of reliability, acceptable generalisability was defined as a coefficient between 0.70 and 0.80, while good generalisability was indicated by coefficients greater than or equal to 0.80 [22]. Figure 2 outlines the design of the Generalisability study and the origin of the variance components.

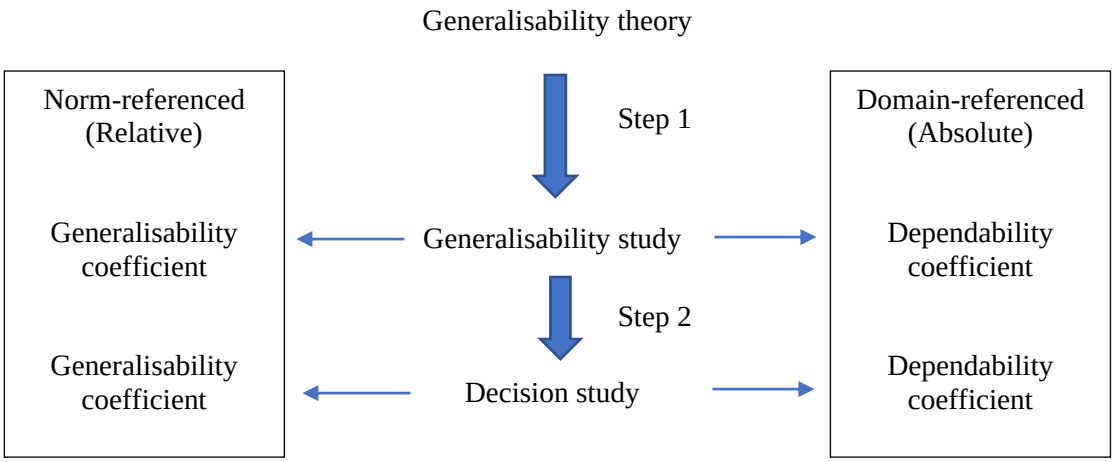

Figure 1: Overview presentation of G-Theory.

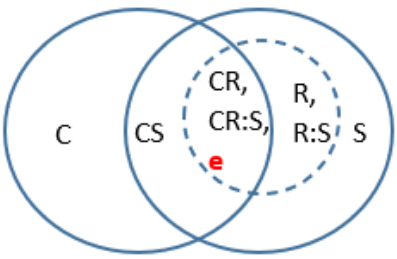

Figure 2: (Raters:Stations) x Candidates (C)

In this study, a nested design was used since each of the ten stations was rated by different groups of three raters. Given that each station comprised an unequal number of attribute scores, a station score was calculated for each candidate by averaging the attribute scores for each rater within a station, resulting in an academic mean score, an administrative means score, and a community mean score for each station. As admission decisions from the MMI were based on rank order by comparing the candidate's performance in each attribute with another candidate, we focused on determining the generalisability coefficient and the relative error variance components. However, our study aims to compare the raters. Therefore, the variance components that estimate error variance for the number of raters and the number of stations in the Decision study investigate how changes to the facets can improve reliability scores. The model underlying the scores for the MMI in our study was a two-facet C × (R:S) random-effects nested design. The random-effects design assumed that the stations and raters were randomly selected from a universe of possible stations and raters.

The formula for the generalisability coefficient in the Generalisability study for this design using a norm-referenced (i.e., relative) approach was:

$$G = \frac{\sigma^2_{(C)}}{\sigma^2_{(C)} + \frac{\sigma^2_{(CS)}}{n_s} + \frac{\sigma^2_{(CR:S)}}{n_s \times n_r}}$$

The formula for the calculation of the dependability coefficient in the Generalisability study using a domain-referenced (i.e. absolute) approach in this design was:

$$D = \frac{\sigma^2_{(C)}}{\sigma^2_{(C)} + \frac{\sigma^2_{(S)}}{n_s} + \frac{\sigma^2_{(R:S)}}{n_s \times n_r} + \frac{\sigma^2_{(CS)}}{n_s} + \frac{\sigma^2_{(CR:S)}}{n_s \times n_r}}$$

Note: Where G = generalisability coefficient, D = dependability coefficient, C = candidates, S = stations, R = raters, R:S = raters within stations, CS = candidate by stations, CR:S = candidate by raters within stations, σ^2 = variance components, n_s = number of stations; n_r = number of raters.

The formulas should provide insight into which sources of information affect the measurement and information used to estimate generalisability coefficients and the differences between absolute and relative estimates. Generalisability and dependability coefficients were obtained separately for each rater. We used these results to inform the Decision study for optimisation of the process.

RESULTS

The mean station scores for each category of raters are presented in Table 1. Statistically significant differences between rater categories were detected for stations 1, 3, 4, 5, 6, and 10. There was a significant difference (p=0.033) between the total mean scores of the academic category (767.7) compared to the administrative category (638.5) and the community category (614.8).

Using the EduG software, the Generalisability study of all raters' variance components of the C × (R:S) random-effects nested design generated the variance components as shown in Table 2. Table 2 shows that the variance component for candidates was 159.35, which accounted for 24.3% of the total variance. The variance component for the candidate showed the difference in how the candidate scored on the overall MMI. It means the extent to which candidates varied in their abilities. The largest variance component was CR:S, candidate score by raters within stations (200.42, 30.5%), which means a large proportion of the effects were caused by interactions between candidates, raters,

stations and other unexplained sources or errors. Even though the lowest variability was seen among stations (19.49, 3%), indicating the station has similar variability, the raters nested within the station (R:S) have the second-highest variability (186.63, 28.4%). This suggests that the raters were not scoring the candidates consistently.

The generalisability coefficient considering all raters together, using a norm-referenced (i.e. relative) approach, was very high (0.91). The variance components applied to the formula that resulted in this coefficient can be verified in the equation below:

$$G = \frac{159.34862}{159.34862 + \left(\frac{90.50897}{10}\right) + \left(\frac{200.41772}{30}\right)}$$

Similarly, using the variance components shown in Table 2, the dependability coefficient considering all raters together, using a domain-referenced (i.e. absolute) approach, was also very high (0.87). The variance components applied to the dependability formula can be verified in the equation below:

$$D = \frac{159.34862}{159.34862 + \left(\frac{19.48717}{10}\right) + \left(\frac{186.62852}{30}\right) + \left(\frac{90.50897}{10}\right) + \left(\frac{200.41772}{30}\right)}$$

The findings of the Generalisability study were used in the Decision study (Table 3). A Decision study helps to project the generalisability that could be expected either through the effect of the number of raters or stations. The Decision Study in Table 3 shows the difference between the generalisability coefficient (relative) and dependability coefficient (absolute). The generalisability coefficient (relative) is calculated considering a norm-reference approach, meaning that scores are only meaningful in relation to each other. The denominator will include the true variance plus the variance components that interact with the person as a result. The dependability coefficient (absolute) is domain-referenced and is interpretable independently, without the need for any comparison with other scores. The denominator includes all variance components and interactions as a result. As our study focuses on just the ranking order of the candidates for admission purposes, we will only interpret the results for the relative generalisability coefficients. Based on the variance components and the random-effects design of ten stations, the generalisability coefficient was 0.85 (one rater), 0.89 (two raters) and 0.91 (three raters). This indicates good reliability for all numbers of raters. The Decision study showed that with one rater, the minimum number of stations needed to achieve a good generalisability was minimal eight stations, while for two raters, five stations were sufficient. If using all three raters together, four stations were sufficient to achieve the desired outcome.

Table 4 shows the Generalisability study for individual raters, a combination of raters, and all raters together. However, under individual reliability, the generalisability coefficient was lowest among the academics (0.76), even though it was still acceptable generalisability and the highest seen in the community (0.91). The administrative and community combination yielded the best generalisability (G=0.94), while the academic and administrative combination produced the lowest but still good generalisability (G=0.88). We ran a simulation by calculating the generalisability coefficient after removing each station and tabulating the stations that contributed to unreliability against each rater category and combination of the rater categories. The result showed that Stations 5, 8, 9, and 10 had significantly impacted reliability.

Table 1: Mean station scores by each rater category

Station	Attributes	Mean station score (standard deviation) by Academic raters	Mean station score (standard deviation) by Administrative raters	Mean station score (standard deviation) by Community raters	*p*-value
1	Conscientiousness	80.9 (7.01)	52.7 (21.9)	53.6 (22.9)	0.002
2	Decision making	49.3 (28.1)	42.2 (29.0)	53.2 (32.3)	0.683
3	Coping with stress	79.9 (15.7)	87.7 (14.7)	54.5 (20.0)	0.001
4	Team player	79.2 (17.0)	85.1 (18.5)	51.9 (21.5)	0.001
5	Altruism	85.6 (8.4)	40.1 (26.6)	72.0 (23.4)	0.001
6	Adaptability	83.6 (16.7)	52.7 (25.7)	55.4 (29.1)	0.010
7	Honesty and integrity	74.0 (27.9)	62.3 (25.6)	65.6 (24.7)	0.562
8	Empathy	92.7 (9.0)	80.9 (25.1)	87.3 (14.9)	0.304
9	Ability to summarise	66.4 (15.8)	65.5 (15.7)	73.6 (17.5)	0.462
10	Active Listening	76.1 (8.8)	69.3 (14.1)	47.7 (24.9)	0.019
	Total Score	767.7	638.5	614.8	0.033

Table 2: Generalisability study of all raters variance component of the C × (R:S) random-effects nested design

Source	σ^2	%
C	159.35	24.3
S	19.49	3.0
R:S	186.63	28.4
CS	90.51	13.8
CR:S	200.42	30.5
G relative (norm-referenced)	0.91	
G absolute (domain-referenced)	0.87	

Notes: C=Candidates (n=11), S=Stations (n=10), R=Raters (n=3), R:S= Raters within stations, CS= Candidate by stations, CR:S= Candidate by raters within stations, σ2 = variance components, % = percentage of variance.

Table 3 : Decision study with generalisability (G-) and dependability (D-) coefficients calculated using the variance components

Number of station	G-Coefficient (norm-referenced)			D-Coefficient (domain-referenced)		
	Only 1 rater	2 raters	All 3 raters	Only 1 rater	2 raters	All 3 raters
3	0.62	0.71	0.75	0.49	0.61	0.67
4	0.69	0.77	0.80	0.56	0.68	0.73
5	0.73	0.81	0.84	0.62	0.72	0.77
7	0.79	0.85	0.88	0.69	0.79	0.82
8	0.81	0.87	0.89	0.72	0.81	0.84
10	0.85	0.89	**0.91**	0.76	0.84	**0.87**
15	0.89	0.93	0.94	0.83	0.89	0.91

Table 4: Generalisability study for a combination of different raters

	G relative (norm-referenced)	G absolute (criterion-referenced)	Stations contributing to unreliability
All raters	0.91	0.87	9
Academic	0.76	0.67	5, 10
Administrative	0.85	0.75	10
Community	0.91	0.87	8, 9
Academic + Aministrative	0.88	0.78	9
Academic + Community	0.91	0.86	9
Administrative + Community	0.94	0.90	None

DISCUSSION

This study investigated differences in candidate performance scores from three different rater categories; academic, administrative and community, using the G-Theory in rating the performance of medical school candidates for admission in an MMI. The psychometric perspective gave an insight into the reliability of the MMI for these three rater categories and with a different number of stations. The key findings were that the three rater categories differed significantly in their total mean station scores. The community was the strictest, and the academic was the most lenient, while the administrative score sat between the community and the academic. Results from the Generalisability study suggested the majority of the observed variance in candidates' scores was due to the raters nested within the station (28.4%), which indicated that the raters were not scoring the candidates consistently.

** Norm-referenced (i.e. relative) refers to scores that have relative meaning (scores have meaning only in relation to each other)

** Domain-referenced (i.e. absolute) refers to scores that have absolute meaning to the domain of measurement

Type of Raters

Our study showed that the community judged the candidates most harshly. This is in contrast with Eva et al. [6], which found that community members scored interviewees marginally higher than academics and Bateman et al. [8], which showed no statistically significant discrepancies between the interview panel's and the community's scores of candidates for admission to health professions training programmes. The perspectives of the community on medical education issues, such as professionalism [3, 23] and communication skills [4, 24], have increasingly been sought. The community had been reported to judge misdemeanours among medical students more harshly than doctors and medical students, implying that their views should be sought when promoting professionalism [23] or indeed, when selecting students for admission to health professions training [7, 8].

One of the most common complaints from the community was the doctor's poor listening skills [25]. Our study demonstrated that the type of rater matters. The result showed a significant score difference between the raters at six of the ten stations where listening skills were among them. While interrater reliability was desirable in all assessments, the homogenous perspective may have misrepresented expectations between the academics, who are the teachers in the medical curriculum, and the community, who are the patients from the community. Including community perspectives and values in student selection ensured that candidates would meet their respective communities' needs while also providing the university with financial and political benefits [26]. Similarly, Eva et al. [27] recommended diversifying the raters across cultures and among subcategories representative of stakeholders within any educational system. This allows raters to draw on their own unique expertise and experience.

Number of Raters

Besides giving an insight into the types of raters, the Generalisability study result also informed the number of raters required to attain good reliability. In our study, the MMI was reliable even with one rater (G-coefficient of 0.85) in a ten-station design. The reliability could be significantly increased by combining the community and the academic raters in the MMI (G-coefficient of 0.94). The G-coefficient for the community alone (0.91), which was higher than academic alone (0.76) as the raters, may suggest that perhaps using a group of trained community raters would be better than using the academic as traditionally practised by most universities. This might free up time for academics to focus on teaching, training, and research, as one of MMI's issues, like the OSCE, is necessary for many raters [28]. The Decision study also found that generalisability was best obtained by increasing the number of raters rather than the number of stations as long as there was rater consistency within stations (Table 3). The Decision Study assists the institution in determining the number of raters or stations required to achieve the target MMI reliability based on existing resources.

Number of Stations

The results from this study provided evidence for the reliability of our MMI framework for admissions. The candidate-station interaction using the MMI instead of a single traditional interview generated more variability among the candidates to provide information for admission decision-making. The MMI's psychometric analysis provided the Malaysian Qualification Agency (MQA) with a justifiable student selection decision. The MQA is the national accrediting body in charge of enforcing course accreditation and regulating the curriculum and operational standards of Malaysia's higher education institutions. The MQA requires that student selection criteria and processes must be clear and consistent with applicable regulations, transparent, and objective [29]. In our study, the

relative G-coefficient of 0.85 for one rater results from the Decision study suggested good reliability of our 10-station MMI. In addition, the result showed that good generalisability could be achieved even with a minimum of eight stations for one rater. The finding is consistent with the literature that while MMI reliability improves with each added station, it tends to plateau at 8–10 stations [27]. This offered us an informed decision to reduce the number of stations needed for the subsequent MMI.

Type of Stations

Our simulation of removing each station and calculating the G-coefficient without the station informed us of the station that needs to be removed or amended. This could indicate a fault with the station's structure, or it could mean that the station's noncognitive attribute was more challenging to assess. The ability to summarise was a noncognitive trait that needs to be re-evaluated in our study to replace it with another crucial noncognitive attribute. If the decision was to keep the station, then the training of the raters should be focused on ensuring that raters understand the definition and expected outcome of this station. Though the rater's training was essential in reducing undesired variance among the raters, the purpose of training is to inform the raters what the MMI is trying to accomplish. The goal is to explore each candidate's unique perspectives rather than focus on eliminating raters' biases. The idea is to benefit from a diverse perspective that occurs when raters' opinions are sampled broadly [27].

Limitations and Directions for Further Research

The limitation of our study was the small sample size. However, the results were favourable for interpretation. This could be attributed to the training given to all rater categories prior to the MMI assessment. The calibration reduced the interrater variance, i.e., inconsistent scores across stations or raters (our "noise"), therefore, allowing a large percentage of the variance to be attributed to the candidates (our "signal"). A new psychometric analysis using the entire cohort of candidates could improve the MMI process. A larger sample could allow the use of the many-facet Rasch model, which could provide a deeper look into the internal structure of the MMI process. Furthermore, as successful candidates go through the medical programme and into practice, a longitudinal research project may be valuable in determining the validity of the admission process.

CONCLUSION

The Generalisability study compared the differences in performance scores from academic, administrative staff, and community members in scoring the performance of medical school candidates for admission to an MMI, shows a significant difference between community members and teachers in judging the noncognitive behaviours. We recommend including community members' perspectives and values in student selection to diversify the raters across cultures and among subcategories representative of stakeholders within any educational system. The high reliability of our MMI framework provided evidence to the accreditation body supporting our student selection procedure. The Decision study informed on the number of stations, the number of raters, and the type of rater combination required to achieve good reliability to provide informed decisions on the process and implementation of the MMI. The station simulation influencing unreliability helps us improve station writing and identify focus areas for training and future station development.

ACKNOWLEDGEMENTS

The authors wish to acknowledge and express appreciation to the MMI pilot Team members (Asso Prof Chen Yu Sui, Asso Prof Verna Kar Mun Lee, Asso Prof Ranjit De Alwis and Mdm Charmaine Khoo) for their contribution in collecting data from running the MMI pilot project. The authors declare that there is no conflict of interest.

REFERENCES

1. Morgan A & Jones D (2009). Perceptions of service user and carer involvement in healthcare education and impact on students' knowledge and practice: A literature review, Medical Teacher, 31(2), 82-95, DOI: 10.1080/01421590802526946
2. Higgins A, Maguire G, Watts M et al. (2011) Service user involvement in mental health practitioner education in Ireland. J Psychiatr Ment Health Nurs 18(6),519–525
3. Jain, A., Petty, E. M., Jaber, R. M., Tackett, S., Purkiss, J., Fitzgerald, J., & White, C. (2014). What is appropriate to post on social media? Ratings from students, faculty members and the public. Medical education, 48(2), 157–169. https://doi.org/10.1111/medu.12282
4. Rimondini, M., Mazzi, M.A., Deveugele, M. et al. (2015). How do national cultures influence lay people's preferences toward doctors' style of communication? A comparison of 35 focus groups from an European cross national research. BMC Public Health 15, 1239. https://doi.org/10.1186/s12889-015-2559-7
5. Poole, P. J., Moriarty, H. J., Wearn, A. M., Wilkinson, T. J., & Weller, J. M. (2009). Medical student selection in New Zealand: looking to the future. The New Zealand medical journal, 122(1306), 88–100.
6. Eva KW, Reiter HI, Rosenfild J, Norman GR (2004). The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. Acad Med 79(6), 602-609.
7. Roberts P, Wild K, Washington K, Mountford C, Caperwell J & Priest H (2010). Inclusion of lay people in the pre-registration selection process. Nursing Standard, 24(48), 42-47.
8. Bateman, H., Smith, M., Melvin, C., Holmes, R. D., & Valentine, R. A. (2019). A Pilot Study to Assess Feasibility of Lay Representation in Dental School Admissions Interviews. Journal of dental education, 83(6), 706–713. https://doi.org/10.21815/JDE.019.077
9. Patrick, L. E., Altmaier, E. M., Kuperman, S., & Ugolini, K. (2001). A structured interview for medical school admission, Phase 1: initial procedures and results. Academic medicine: journal of the Association of American Medical Colleges, 76(1), 66–71. https://doi.org/10.1097/00001888-200101000-00018
10. Kreiter, C.D., Yin, P., Solow, C. et al. (2004). Investigating the Reliability of the Medical School Admissions Interview. Adv Health Sci Educ Theory Pract, 9, 147–159 https://doi.org/10.1023/B:AHSE.0000027464.22411.0f
11. van der Vleuten, C. P. (2014). When I say … context specificity. Medical Education, 48(3), 234-235.
12. Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. Medical education, 38(3), 314–326. https://doi.org/10.1046/j.1365-2923.2004.01776.x
13. Pau, A., Chen, Y. S., Lee, V. K., Sow, C. F., & De Alwis, R. (2016). What does the multiple mini interview have to offer over the panel interview?. Medical education online, 21, 29874. https://doi.org/10.3402/meo.v21.29874
14. Norman, G. R., Van der Vleuten, C. P., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. Medical education, 25(2), 119–126. https://doi.org/10.1111/j.1365-2923.1991.tb00037.x
15. Pau, A., Jeevaratnam, K., Chen, Y. S., Fall, A. A., Khoo, C., & Nadarajah, V. D. (2013). The Multiple Mini-Interview (MMI) for student selection in health professions training - a systematic review. Medical teacher, 35(12), 1027–1041. https://doi.org/10.3109/0142159X.2013.829912

16. Cardinet, J., Johnson, S., & Pini, G. (2011). Applying generalizability theory using EduG.
17. IRDP (2016). Generalizability Study. Working Group - Edumetrics - Quality of measurement in education of the Swiss Society for Research in Education. Retrieved on 01 September 2022 at https://www.irdp.ch/institut/english-program-1968.html
18. Sebok SS, Luu K and Klinger DA. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. Adv in Health Sci Educ, 19, 71-84
19. Shavelson RJ and Webb NM (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publication
20. Brennan, R.L. (1992), Generalizability Theory. Educational Measurement: Issues and Practice, 11, 27-34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x
21. Lambert W. T. Schuwirth & Cees P. M. van der Vleuten (2011) General overview of the theories used in assessment: AMEE Guide No. 57, Medical Teacher, 33(10), 783-797, DOI: 10.3109/0142159X.2011.611022
22. Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. Lancet, 357, 945-949.
23. Brockbank, S., David, T. J., & Patel, L. (2011). Unprofessional behaviour in medical students: a questionnaire-based pilot study comparing perceptions of the public with medical students and doctors. Medical teacher, 33(9), e501–e508. https://doi.org/10.3109/0142159X.2011.599450
24. Bensing JM, Deveugele M, Moretti F, et al. How to make the medical consultation more successful from a patient's perspective? Tips for doctors and patients from lay people in the United Kingdom, Italy, Belgium and the Netherlands. Patient Educ Couns. 2011;84(3):287-293.
25. Kee, JWY, Khoo HS, Lim I, Koh MYH. (2018). Communication Skills in Patient-Doctor Interactions: Learning from Patient Complaints. Health Professions Education, 4(2), 97-106. https://doi.org/10.1016/j.hpe.2017.03.006.
26. Stagg, P., & Rosenthal, D. R. (2012). Why community members want to participate in the selection of students into medical school. Rural and remote health, 12, 1954.
27. Eva, K. W., Macala, C., & Fleming, B. (2019). Twelve tips for constructing a multiple mini-interview. Medical teacher, 41(5), 510–516. https://doi.org/10.1080/0142159X.2018.1429586
28. 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), e1437–e1446. https://doi.org/10.3109/0142159X.2013.818634
29. Malaysian Qualifications Agency (2017). Student Selection: Code of Practice for Programme Accreditation. 2nd Edition, page 16. Retrieved on 06 February 2022 at https://www2.mqa.gov.my/qad/garispanduan/COPPA/COPPA 2nd Edition (2017).pdf