**ORIGINAL ARTICLE**

# INTAKE INTERVIEW PROCEDURE FOR THE SELECTION OF UNDERGRADUATE STUDENTS IN MALAYSIAN MEDICAL SCHOOLS: INTER-RATTER RELIABILITY OF QUALITY ASSESSMENT

**[1]Shahid Hassan, [2]Nordin Simbak, [3]Harmy bin MohdYussof, [4]Myat Moe ThweAung**

[1]Unit of Medical Education, [2]Unit of Orthopaedic, [3]Unit of Family Medicine, [4]Unit Community Medicine, Faculty of Medicine Universiti Sultan ZainalAbidin, Malaysia

## ABSTRACT

*Comparable selection methods based on interview as one of the selection criteria are used in many countries globally however; procedure of interview and its reliability has been of varying nature. A semi-structured interview procedure was developed by the Faculty of Medicine at Universiti Sultan ZainalAbidin to finally select the shortlisted candidates seeking to studying medicine in this institution as the new intake of 2015-2016 sessions of MBBS program. Multiple panels comprising of two members each to independently select the candidate held interview. Inter-ratter reliability of quality assessment was investigated. Current article investigates the inter-ratter reliability of interviewers in quality assessment of candidates seeking to join the Faculty of Medicine at Universiti Sultan ZainalAbidin, Malaysia. An observational study, conducted across all the candidates, who were shortlisted on merit for formal selection through interview procedure. Data reflecting candidates' characteristics and qualities were collected as quantitative score. Inter-ratter reliability using intra class coefficient was calculated for interpretation. A moderate difference of mean (SD) among the interviewer varying from 37.61 (3.48) to 42.12 (0.60) was observed. The reliability of score varied between 0.50- 0.65, significant at p = < 0.05 with majority assessors. However, among the 4 panels of assessors' intra-class correlation coefficient was between 0.70-0.0.90 (p = < 0.001). Assessment of candidates' performance based on observation did not achieve the satisfactory level of intraclass correlation coefficient (ICC ≥ 0.70). However for higher discrepancy in inter-ratter scores in some cases, continuing faculty development program in interviewing skills and calibration workshops are recommended to improve the reliability and validity of quality selection through interview procedure in future.*

***Keywords**: inter-ratter reliability, reliability coefficient, structured interview procedure, interviewing skills.*

## INTRODUCTION

Selection of students for undergraduate medical education (MBBS and MD) in Malaysia has been the core business of Ministry of Education selection is based on candidate's merit in prequalification examination (Matriculation, Foundation and STPM) results. This is done at national level in which candidates are asked to give their priority of choices of ranking order to join a medical school. USM being the APEX University was the first medical school in public sector to be allowed to shortlist (based on their prequalification national merit) and call the candidates for an interview before selecting them to join MD program in that school. Currently, since 2015 all public universities are allowed to interview candidates on merit besides English proficiency test (minimum band 3). Faculty of Medicine in Universiti Sultan ZainalAbidin (UniSZA) developed their interview procedure and faculty to polish their interviewing skill. The first intake based on merit plus interview procedure was held in May 2015 by a trained panel of 12 interviewers in three different centres all over Malaysia.

Comparable selection methods based on interview as one of the selection criteria are used in many countries globally[1] however; procedure of interview and its reliability has been of varying nature. Inter-ratter reliability is often not evaluated in many instances. In general, the reliability of interview assessments in medical school admission is considered moderate to good[2]. Reliability increases by structuring interviews, training assessors and increasing the number of assessors or interviews[3,4,5]. Interview assessment method has never been satisfactory for many reasons including structure of interview procedure, number of instrument used, training of the assessors and their credibility, calibration process and subjective bias of assessors, influence of hierarchy and politically motivated forces in communities. There is not a single interview method that can ensure the right selection to predict competent doctors in future[6]. Based on current scenario associated with many suspected bias in interview methods and candidates' fair selection, Faculty of Medicine developed their first manual of interview protocol with as many as possible components, standardized questions and scenarios to mark candidates basic competencies required for studying medicine. Motivation and general knowledge, observable personal attributes and communication skills, language proficiency in BahasaMaleyu and English languages and co-curricular activities at district, national and international levels (see appendix) are components to judge candidates' on merit.

Reliability studies are widely used to assess the measurement reproducibility of human observers

when it is randomly repeated for the same subjects, particularly using interview as method of selection. Measurement reliability plays a very important role, since it affects the choice of the primary outcome. Usually, reliability studies are conducted to assess the measurement of reproducibility of tests and coefficient of variation is used. However, it has been demonstrated that a coefficient of variation does not measure reliability[7]. Measurement of reliability particularly associated with interview methods in candidate's selection inter-ratter agreement becomes more important to achieve reliability. It is a truism of arbitration that the process is only as good as the quality of the arbitrators conducting it[8]. It is also a truism that an institution will strive to select an arbitrator who has some inclination or predisposition to favour the institution policies and likely decisions[9]. Provided that the arbitrator does not allow personal bias or compromise on professional judgment, we tried to minimise such influences biases by keeping faculty members involved in structuring of interview procedures and marking schemes by their training through hands-on workshop to polish their interviewing skills, independent evaluation using a checklist as well as global rating with remarks in their final selection minimising bias as much as possible.

As we decided to maintain a highly structured interview in the new procedure, we investigated the reliability of interview assessments for its inter-ratter reliability in the current selection procedure that interview is one of the selection criteria. For quantitative measures, intra-class correlation coefficient (ICC) is considered principal measurement of reliability as one of the best measure of reliability using a continuous data and intra-class correlation coefficient[10]. We used ICC to interpret the inter-ratter reliability of assessors involved in interview procedure in current study.

## METHODOLOGY

### Design

An observational study of candidates who are initially selected by Ministry of Education based on following criteria.

1) STPM (Higher Secondary Board Examination) or Matriculation examination result with minimum CGPA 3.60. 2) Minimum grade B with essential subjects of biology, chemistry, physics and maths in STPM and Matriculation and advance maths in STPM examination. 3) Band 3 and above in MUET (English language proficiency test). 4) Physically and mentally healthy.

Those selected through the above mentioned criteria were invited by administration to register for interview for one of the three venues in three different cities in Malaysia. Interview panel was selected from those faculty members who received training through a course and workshop organised for their training in interviewing skills using newly developed interview procedure. 1-day workshop mainly focused on introduction of interview manual, standardized questions, marking scheme and development of various scenarios to be used to judge candidates personal attributes and communication skills as one of the major component. Each interviewer was advised to assess candidates' quality and personal attributes (see appendix) independently to avoid each other influence on selection of a candidate. Personal interview was scheduled for 25-30 minutes using two languages (BahasaMelayu and English) in order to mark their proficiency in these two languages. In a semi-structured interview same set of two scenarios were ultimately used by all the members of panel. Characteristics observed were the candidates' motivation to study medicine, their general knowledge in relation to current health or environmental community or global issue, language proficiency, personal attributes, communication skills and co-curriculum engagement at district, national or international levels.

### Data Collection

Prescribed format marking scheme was used to document students' performance in interview session, besides demographic information regarding the name, gender and sex and their academic and co-curricular activities provided to candidates. The candidate's characteristics and quality was independently recorded on a 5-point likkert scale by two members of the interviewing pane. An average of two scores was the candidates' ultimate score for further analysis.

### Data Analysis

SPSS version 22 was used to analyse the data. The reliability aspects were described with Mean Score (SD) according to multiple panels of paired assessors (see table 1). Inter-ratter reliability was established for each score with intra-class correlation coefficient (ICC) calculated for correlation of two assessors in each panel (see table 2). F statistics for significance of inter-ratter agreement was noted.

## RESULT

In descriptive statistics slight to moderate difference of Mean Score (SD) were observed between the paired assessors of multiple panels (see table 1) however, small SD with most of the panels were indicative of nearly normal distribution. The inter-ratter reliability of scores among assessors was generally on the lower side indicating more towards disagreement and the acceptable range 0.72 and above was achieved by 4 (33.33%) panels only However, majority panels 7 (58.33%) were between > 0.5 - < 0.7. One panel each (8.33) was excellent (> 0.90) and poor (< 3.50) was excellent (> 0.90) to poor (< 0.50) among the overall 12 panels involved in interview process, among few

ratters with 95% confidence interval. F statistics however, was significant < 0.05 whereas two of the panels F test was insignificant (*P* > 0.05) between the two ratters involved in interview process (see table 2)

**Table 1: Descriptive statistics of all 16 individual assessors' performance (some assessors were repeated and performed with changing partners)**

| No | Frequency (%) | Mean (SD) | Minimum | Maximum | Skewness | Kurtosis |
|---|---|---|---|---|---|---|
| 1 | 10 (5.7) | 45.33 (2.83) | 39.7 | 49.0 | -0.673 | 0.239 |
| 2 | 41 (23.6) | 43.06 (4.55) | 30.5 | 49.0 | -0.905 | 0.444 |
| 3 | 47 (27.0) | 41.13 (4.43) | 31.3 | 50.0 | -0.392 | -0.564 |
| 4 | 22 (12.6) | 40.48 (4.97) | 33.5 | 48.2 | -0.029 | -0.790 |
| 5 | 36 (20.7) | 43.50 (3.83) | 33.3 | 50.0 | -0.953 | 1.358 |
| 6 | 45 (25.9) | 44.14 (4.02) | 33.3 | 50.0 | -0.895 | 0.886 |
| 7 | 30 (17.2) | 41.12 (4.52) | 30.5 | 50.0 | -0.490, | -0.106 |
| 8 | 18 (10.3) | 41.48 (4.41) | 33.0 | 47.7 | -0.619 | -0.509 |
| 9 | 28 (16.1) | 37.61 (3.48) | 29.9 | 46.7 | -0.026 | 1.673 |
| 10 | 25 (14.4) | 40.03 (5.41) | 29.0 | 49.0 | 0.234, | -0.347 |
| 11 | 9 (5.2) | 46.70 (3.92) | 37.0 | 50.0 | -0.2.04 | 4.99 |
| 12 | 9 (5.2) | 37.88 (4.17) | 34.2 | 46.7 | 0.127 | 0.124 |
| 13 | 8 (4.6) | 47.22 (1.62) | 44.3 | 49.0 | -0.763 | -0.066 |
| 14 | 6 (3.4) | 48.11 (.598) | 47.50 | 49.0 | 0.768 | -1.273 |
| 15 | 7 ((4.0) | 43.61 (5.78) | 34.2 | 50.0 | -0.991 | -0.472 |
| 16 | 7 (4.0) | 45.22 (3.26) | 39.0 | 48.7 | -1.26 | 1.63 |

**Table 2: Intra-class Correlation Coefficient as inter-ratter reliability of 16 Assessors in 12 panels (some of the assessors were repeated in more than one panel where as two panels were repeated as such with same assessor) in Faculty of Medicine at UniSZA.**

| Panel No | Intra-class Correlation Average Measure | 95% Confidence Interval | | F Test | | |
|---|---|---|---|---|---|---|
| | | Upper Bound | Lower Bound | Value | df. | Sig. |
| **1** | .909 | .653 | .977 | 12.195 | 9 | .000 |
| **2** | .827 | .354 | .940 | 8.767 | 21 | .000 |
| **3** | .667 | .351 | .830 | 3.004 | 35 | .001 |
| **4** | .826 | .316 | .940 | 8.961 | 22 | .000 |
| **5** | .530 | -.280 | .825 | 2.091 | 17 | .069 |
| **6** | .585 | .006 | .835 | 2.776 | 18 | .018 |
| **7** | .742 | .026 | .940 | 4.482 | 8 | .024 |
| **8** | .619 | -.288 | .909 | 4.356 | 8 | .026 |
| **9** | .569 | -.472 | .905 | 2.744 | 7 | .103 |
| **10** | .331 | -1.854 | .897 | 1.569 | 5 | .317 |
| **11** | .981 | -.880 | .997 | 66.791 | 6 | .000 |
| **12** | .594 | -.342 | .921 | 4.513 | 6 | .045 |

**DISCUSSION:**

The mean score (SD) showed some variation of qualities, reflecting the candidates performance and it can be attributed to ratters inherited stringy vs. lenient approach in marking, problem of active participation in faculty development training and workshop and issues related to calibration of inter-ratter agreement[11]. However, the selection criteria, duration of interview time and random selection of interviewing panel were reported generally satisfactory. This might have been due to design of semi-structured protocol of interview procedure with addition of multiple components, logically attributed marks according to importance and weighting of each component, assessment of personal attributes using two instead of one and same instead of different scenarios to standardize

the observational judgement and random rotation of the panel members.

Reliability refers to consistency as well as inter-ratter agreement of score given by the assessors in an interview procedure used to select quality students for any given program. ICC can calculate consistency as well as inter-ratter agreement. Difference of inter-ratter agreement suggest that the both ratters are not observing candidates in similar way and one ratter may provide low rating while the other ratter may provide high rating. This may result in discrepancy of absolute agreement however; it is possible for consistency to be high if the rank ordering of these rating were similar. In current study reliability in terms of inter-ratter agreement using ICC statistical test is explored.

Reviews of published literature have shown varying reliability perhaps due to widely varied format, structure of interview procedures and subjective bias among the assessors[11,12,13,14]. The current study demonstrates less than the expected satisfactory level of reliability of candidate's quality assessment, which is not comparable with some of the recent studies[15,16]. Factors known to enhance the reliability depend on effects of structure of interview procedure, training of interviewers and sharing of previous experience[14,17,18]. Current study also endorse the idea of engaging 2 assessors per panel and multiple but trained panellists, optimal interview time (30 minutes) as it has been found the reliability of an assessment procedure partly depends on the duration of the procedure[19]. Two or more standardized scenarios with marking guidelines for all candidates and a semi-structured interview procedure comprising of optimum numbers of components and a well-designed marking scheme was used. The gender of assessors' bias was not a factor in reducing the reliability since most of the panels were of mixed pair of male and female faculty members

With regards to strengths and weaknesses of current study, idea of conducting the interview with multiple panels and venues did not burden the assessors. Random selection and changing the order of pair in case of repeatedly involved assessors was a good experience to keep up the motivation of interviewers. Besides, multiple components (6 components) and multiple attributes of each component were considered a strong point of the current study. Two assessors per panel was though helpful but it did not managed to achieve the desirable level of inter-ratter agreement between the assessors and this is contrary to studies that find satisfactory reliability between 2 assessors[20,21]. Lack of previous interviewing experience and data was among the weak factors of current study. A 5-point likert scale was considered a limited choice rubric to interfere with inter-ratter reliability felt by some assessors after experiencing the interview procedure and a likert scale with 9-10 points' judgement was desired. However, literature suggest that reliability of rating at the higher or lower end of rating scales is higher than that of middle levels and even a three-point scale has been found as useful as the commonly used five-point scale[22,23].

**CONCLUSION:**

Assessment of candidates' performance with high level of agreement among the assessor based on observational interview of new intake was though not adequately satisfactory to achieve the required level of inter-ratter reliability was a good experience with evaluation of entire process. Calibration workshop to improve the inter-ratter reliability in future is recommended. However, high discrepancy in inter-ratter scores to determine the quality and characteristics of candidates were hardly observed. Minor to moderate level of disagreement between the assessors in in general is attributed to problems of adequacy of time available for faculty development in interviewing skills, and hands-on calibration exercises. Adequate training and calibration workshop is therefore recommended to improve the reliability and validity of a competence based interview procedure with required level of inter-ratter agreement among the assessors.

**ACKNOWLEDGEMENT:**

The authors would like to thank all the faculty members inclusive of Prof. Dr Wan Pauzi, Prof. Dr Zawawi bin Nordin, Prof. Abdul Mutalib, Prof. Asmabinti Hassan, Prof. Dr Tengku Mohammad Arif, Associate Prof. Dr RahmahMohd. Amin, Associate Prof. Shamim Ahmed Khan, Associate Prof. AnizabtAbd. Aziz, Associate Prof. Dr Tg Fatima Murniwati, Associate Prof. Dr Izabt A. Rahman, Dr.Norhasizahbt Mat Jusoh, and staff of the Academic Office including Ms SazmawatiShamsuddin, Ms BaheyahSanat, and last but not the least staff of administrative office Mr.ZulkifliAwang and Ms.RasmahEmbong.

**REFERENCES:**

1 Vermeulen MI, Kuyvenhoven MM, Zuithoff NPA, Tromp F, van der Graaf Y, Pieters HM: Selection for Dutch postgraduate GP training; time for improvement. Eur J Gen Pract 2012, 18:201–205.

2 Kreiter CD, Yin P, Solow C, Brennan RL: Investigating the reliability of the medical school admissions interviews. Adv in Health SciEduc 2004, 9:147–159.

3 Albanese M, Snow M, Skochelak S, Huggett K, Farrell P: Assessing personal qualities in medical school admissions. Acad Med 2003, 78:313–321.

4 Salvatori P: Reliability and validity of admissions tools used to select students for the health professions. Adv in Health SciEduc 2001, 6:159–175.

5 Morris JG: The value and role of the interview in the student admission process: a review. Med Teach 1999, 21:473–481.

6 Siu E, Reiter HI: Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. Adv in Health SciEduc 2009, 14:758–775

7 John M Lachin (2004). The role of measurement reliability in clinical trials. Clinical trials 2004, 1: 553-566.

8 Hunter &Paulsson, A Code of Ethics for Arbitrators in International Commercial Arbitration?, 19 Int'l Bus. Law 153, 155 (1985).

9 A. Redfern & M. Hunter, Law and practice of international commercial arbitration 217 (2d ed. 1991).

10 Margit I Vermeulen, Marijke M Kuyvenhoven, Nicolaas P A Zuithoff, Yolanda van der Graaf and Roger A M J Damoiseaux. Dutch postgraduate GP selection procedure; reliability of interview assessments. Vermeulen et al. BMC Family Practice 2013, 14:43 http://www.biomedcentral.com/1471-2296/14/43.

11 *MuhamadSaifulBahriYusoff, Ahmad Fuad Abdul Rahim*. The discrepancy-agreement grade (DAG): A novel grading system to provide feedback on rater judgments 2012, Vol 4, No 2.

12 Kreiter CD, Yin P, Solow C, Brennan RL: Investigating the reliability of the medical school admissions interviews. Adv in Health SciEduc 2004, 9:147–159.

13 Salvatori P: Reliability and validity of admissions tools used to select students for the health professions. Adv in Health SciEduc 2001, 6:159–175.

14 Edwards JC, Johnson EK, Molidor JB: The interview in the admission process. Acad Med 1990, 65:167–177.

15 Lumb AB, Homer M, Miller A: Equity in interviews: do personal characteristics impact on admission interview scores? Med Educ 2010, 44:1077–1083.

16 Rao R: The structured clinically relevant interview for psychiatrist in training (SCRIPT): a new standardized assessment tool for recruitment in the UK. Acad Psychiatry 2007, 31:443–446.

17 Albanese M, Snow M, Skochelak S, Huggett K, Farrell P: Assessing personal qualities in medical school admissions. Acad Med 2003, 78:313–321.

18 Morris JG: The value and role of the interview in the student admission process: a review. Med Teach 1999, 21:473–481.

19 van der Vleuten CPM, Schuwirth LW: Assessing professional competence: from methods to programmes. Med Educ 2005, 39:309–317.

20 Hamel P, Boisjoly H, Corriveau C, Fallaha N, Lahoud S, Luneau K, Olivier S, Rouleau J, Toffoli D: Using the CanMEDS roles when interviewing for an ophthalmology residency program. Can J Ophthalmol 2007, 42:299–304.

21 Patrick LE, Altmaier EM, Kuperman S, Ugolini K: A structure interview for medical school admissions, phase 1: initial procedure and results. Acad Med 2001, 76:66–71.

22 Prideaux D, Roberts C, Eva K, Centeno A, McCrorie P, McManus C, Patterson F, Powis D, Tekian A, Wilkinson D: Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 conference. Med Teach 2011, 33:215–223.

23 Stansfield R, Kreiter C: Conditional reliability of ratings: extreme ratings are the most informative. Med Educ 2007, 41:32–38. 25. Siegel S, Castellan NJ: Nonparametric statistics. 2nd edition. New York, USA: New York McGraw-Hill Book Company; 1988:262–272.