ORIGINAL RESEARCH

Qualitative Analysis of Multiple Mini Interview Interviewer Comments

R. Stephen Manuel[1] · Lesley Dickens[1] · Kathleen Young[1]

Abstract

Objective Qualitative studies of the Multiple Mini Interview (MMI) have investigated the attitudes and thoughts of prospective students and interviewers (i.e., raters) on the MMI interview, but none have examined rater's written assessments. Concerns regarding what the MMI measures, especially across and within each interview, have sparked investigations to determine how and what raters are measuring. Raters communicate their student evaluation(s) through numerical ratings and written comments that provide score context. This study explores rater's written comments to better understand the specific information gathered during the MMI process that contributes to interviewee evaluations.

Methods Randomized data from two US medical schools were examined with no numerical scores or other information about the interviewee provided to reviewers. In reviewing the rater comments, common words and phrases were identified to help construct themes that characterized the content (domains). Authors reviewed each other's notes and comments regarding themes and worked together to verify themes for accuracy.

Results Using a directed content approach to content analysis and reviewing the rater's comments, the results indicate that raters are focused on seven different domains: perspective taking, presentation, qualities, communication, coherence, comprehension, and non-verbal. Many of the rater comments contained multiple themes.

Conclusion Raters' MMI comments provide the context for numerical scores allowing admissions committees to more fully understand a candidate's strengths or weaknesses. Identifying the themes in rater comments can ultimately assist the admissions committee to more comprehensively understand assessment elements that raters are using and consider important during the MMI evaluation.

Keywords Admissions · Multiple Mini Interviews · Qualitative

Introduction

The Objective Structured Clinical Examination (OSCE) was born out of frustration with prior assessment methods and sought to bring both reliability and validity to the clinical assessment of medical students [1]. Prior to the development of the OSCE, students were assessed clinically based on different cases, patients, but the assessments were not reliable [2]. By standardizing the clinical assessment to a standardized patient, reliability increased and in theory, subjectivity was minimized. During the

OSCE, each patient is referred to as a "station" and the students are assessed based on a scoring rubric. The number of stations that an OSCE has increases the reliability. Research has demonstrated that while there might be low generalizability within and across different stations and domains, global scores can show moderate reliability [3]. The variance in the student assessment score is affected by the reviewer, the patient and the illness that the patient reported to be acting. Even in reviewing a taped OSCE, two assessors can reach a different score; hence, why the OSCE is performed multiple times.

In a similar fashion and for the same reasons that the OSCE was created, the Multiple Mini Interview (MMI) was developed to replace the medical school admissions interview [4]. The MMI is an interviewing technique used by medical and other professional schools to assess communication skills in prospective students and is based on the same format as the OSCE [5, 6]. The MMI is normally

✉ R. Stephen Manuel
    manuelrsteve@gmail.com

1   University of Mississippi Medical Center, 2500 North State St., Jackson, MS 39216, USA

multiple short (mini) interviews that are focused on a scenario and is behaviorally based. The MMI has been shown to predict clinical performance during the third year and clinical exam grades (Eva, 2004 et al). Qualitative studies have been conducted to determine attitudes and thoughts of prospective students concerning the MMI [7] and with interviewers (referred to as raters) [8], but to our knowledge, no qualitative research has examined written comments submitted by MMI interviewers. This area is of critical importance because questions have been raised concerning what exactly the MMI measures, especially across and within each interview [9, 10]. These questions have led to investigations to determine how and what raters are actually measuring in their scores [11].

As an example, one of the strengths of the MMI is that using multi interviews (stations), generally 8–10, yields scores with greater reliability in comparison with the traditional, one-on-one interviews [12]. The reliability within stations is typically low; however, the reliability across stations is higher ([6]. In layman's terms, this means that if raters are assessing multiple traits during an MMI (sub-scores such as empathy, interest in scenario, and verbal ability), the individual sub-scores in each of these traits are inconsistent across different stations (for instance, sub-scores for evaluating empathy across multiple stations are not reliable (Zaidi, Swoboda, Wang, & Manuel, 2014)). However, when asked to give an overall or global score, the overall score is reliable across stations and scenarios. Why are raters in agreement regarding the overall score?

Some research suggests that, instead of assessing separate attributes, the MMI is assessing one unidimensional attribute [9]. This unidimensional attribute theory is supported in original "Experiment 1" in the development of the MMI. In the original development of the MMI instead of evaluating prospective students in four areas (communication, strength of arguments raised, suitability for the health sciences, and overall performance), these areas were combined during the pilot to a single overall score due to the "high correlations among the 4 evaluation questions" ([4], p. 318). This single trait concept theory is not unique to the MMI and is also found for the OSCE. Interestingly, several papers have suggested that only "global" scores should be used for the OSCE due to increased psychometric properties [13].

The present study attempts to fill the gap in the research concerning rater assessment and to determine what raters might be assessing with their global score. Since the global score is reliable across stations, the study seeks to determine if there are agreed-upon traits that all raters are using to generate the global score. This study is a qualitative assessment of the MMI interviewer comments. Through this examination, this study will elicit the attributes that the raters think are important in evaluating the prospective student during the MMI and thus give valuable insight into what is being actually assessed during the MMI interview.

Method

The study was conducted using data from two US medical schools (noted as A and B), and institutional review board approval was obtained at both schools. School A is located in the Midwest and School B is located in the South. Both schools have used the MMI for over 5 years and utilized the same software system to record overall score and rater comments on the assessment of prospective students. In addition to using ProFitHR (www.profithr.com) training materials, both schools use the same videos and similar manuals to train raters. The ProFitHR training materials include an outline of the MMI, information for interviewers on the purpose and use of the MMI, and example scenarios.

The initial training at both schools lasts approximately 90–120 min. During the training session at both schools, 3 examples of an MMI are scored and the scoring methodology is reviewed. The length of the MMI comments and how the comment is used by the committee in reviewing the MMI performance are also explained and demonstrated. Both schools give examples of comments that are viewed as useful and those that are unsatisfactory in terms of language and length. The raters are specifically advised to "provide context for your score" through the comment section. Raters are told that comments should inform the committee on why the prospective student is receiving the score being given to them [14]. The raters are informed that the committee is looking for patterns that span multiple MMI stations and comments provide examples of consistent student behavior.

A circuit represents the number of stations (normally scenarios) that the prospective student completes. At one medical school, the circuit consisted of ten stations (school B), at the other eight (school A). Rater open-ended comments were randomly generated using one years' worth of comments (fall 2014 School A, and fall 2015 school B). Both schools listed sub-score sections on the student evaluation form that were identical (nonverbal, interpersonal, articulation, interest in dilemma, reflection, multiple perspectives) and that are designed to be used as a guide for the overall score. At both schools, the raters enter their scores and comments into a software system. The system allows the review of all scores and comments for each prospective student for all interviewers in one table.

Data Analysis

At each school, there are over 250 students who participated in the MMI. One school interviewed prospective students with 8 stations, yielding over 2000 (8 × 250) scores and comments and the other school had over 2500 scores and comments. Data for the MMI scenarios, student id, comments, and scores were downloaded into two separate databases, one for each school. The data was reviewed to determine that each school was using different MMI scenarios, and then the data was randomized by scenarios, students, and scores. Once randomized, the student's name, MMI score, and any additional information other than the comment were removed.

Using conventional content analysis [15] and a directed approach, two authors (A and B) independently reviewed the open-ended comments made by raters in assessing the student performance. Authors were provided the student ID and rater comments. No additional information concerning the students or raters was provided. Author A and author B read between 150 and 300 random comments from raters at one of the two schools, author A from school A and author B from school B. The difference in the number of prospective students reviewed is due to the larger number of prospective students interviewed at school B. In both cases, the comments were not only random, but each comment was regarding a different student. Both reviewers were given the information on an excel spreadsheet.

Each author initially read through the rater comments repeatedly to obtain a global perspective and identify themes or pattern. The authors were looking for common words, phrases, and domains (a summary of themes, patterns, or classifications) that characterize phrases [16] and made notes next to rater comments. Taking notes for the authors meant highlighting words, putting common words in a separate column, or making comments on possible domains in a separate column. Both authors kept word counts for common words as a tool and each developed domains independently. Authors reviewed the list of comments and their own notes multiple times over several weeks to determine domains. Once the authors had completed their review, they submitted their initial identified domains.

Taken separately, there were 13 domains that the authors identified independently (with considerable overlap). Authors A and B exchanged the initially identified domains prior to the first review meeting and had an opportunity to review the new combined list and their own comments again. All of the authors then met to review the common words, phrases, and domains; reconcile differences; and discuss the common emerging domains. After discussing and reviewing the domains and referring to the comment database (both schools), the three authors agreed on the overall domains. Seven domains emerged from the rater comments and three sub-themes were identified across three of the major domains. See Table 1.

Table 1 The seven domains and the three sub-themes

Domains	Sub-themes
Perspective taking	
Qualities	Character
Coherence	Clarity
Communication	
Comprehension	
Non-verbal	Body language
Presentation	

Sub-themes are areas within the domains that represented one of the most cited examples. For example, body language was one of the most cited non-verbal areas mentioned. After identifying these sub-themes, the authors had the opportunity to review them and suggest any changes. No changes were suggested.

Many of the rater comments contained multiple domains (e.g., non-verbal, qualities): "she had good body language and it was easy to see that she cared, but her speech was slow and labored and she seemed flustered." In this case, it appears that non-verbal behavior was a key in the assessment of the student performance, but the rater is also adding an additional quality that the rater believes can be assessed: caring.

Raters mentioned many different types of qualities that they believed could be assessed during the 8-min encounter (italics by authors):

> "*Thoughtful* and *humorful* candidate. Mentioned the importance of breaking habits, establishing a healthier lifestyle through patient education, advocacy, support groups, and increases in self-efficacy. Expressed *sensitivity* toward patients of varying backgrounds, as well as concern for not penalizing high risk patients with higher healthcare costs due to the economic hardships it would cause them. Seemed like a *kind-hearted* and *down-to-earth* applicant."

There are other raters who appeared to stay much closer to the sub-scores and actually used the sub-scores in the written evaluation of the candidate (e.g., perspective taking, communication, non-verbal, presentation, coherence (clarity)):

> "… was very *respectful* and professional when discussing this scenario. He frequently mentioned the *multiple perspectives* and viewpoints that are at play here, and how these *perspectives* can lead people to believing certain things. He was very *clear*, *concise*, and *detailed*, and gave well-thought conclusive answers. Furthermore, he made very good use of *nonverbal* and *interpersonal skills* when delivering his message, and he demonstrated a sincere interest in the issue."

Some raters were able to touch on a number of domains and write objectively about the candidate's performance across each domain:

(perspective taking, non-verbal (body language), coherence)

"This applicant considered limited *perspectives*, even with extensive prompting. She did not elaborate greatly on the scenarios and questions presented to her. She had good *eye contact* and was quite polite, but did not demonstrate *enthusiasm* or *confidence*. Her *body language* was withdrawn and she frequently wavered and was uncertain of her own opinion."

(qualities, communication, non-verbal)

"Super *dynamic*, *high energy. Excited.* Talked a lot about a positive attitude dealing with high expectation. *Articulates* well, very *personable* and easy to talk to. Excellent *tone, volume, eye contact*"

(communication, non-verbal, perspective taking, qualities)

"*VERY awkward*. Did not preface the conversation at all, just sat down and started talking about stopping smoking. Very weird *tone inflection, changing pitches* at *awkward times*. Minimal *eye contact*, lacked any productive suggestions, seemed *closed minded* and did not show much *empathy*."

(coherence, presentation, perspective taking)

"So *nervous* it was uncomfortable to speak to him. Paused multiple times to collect his thoughts. Required much probing, though answers remained shallow. Focused more on the laws than on the ethical nature of the scenario."

These are examples of comments that support the seven domains and three sub-themes that the authors identified. There were many comments that were very brief and others that were very long. The examples given should not be interpreted in terms of comment length. It should be noted that there were comments that did not include any of the identified themes.

Discussion and Conclusion

In this study, we used MMI rater comments from two schools to determine what raters perceived to be important to the admissions committee in describing how the prospective students performed. Using a directed approach to content analysis and reviewing the rater's comments, the results that emerged indicate raters are focused on seven different domains: perspective taking, presentation, qualities, communication, coherence, comprehension, and non-verbal.

The authors could find no qualitative or quantitative assessment of the rater comments from the MMI in the literature. However, the MMI was modeled after the OSCE and similar studies on the OSCE have attempted to assess what interviewers are actually assessing [17]. Psychometric issues in assessment for the OSCE that have been identified have also now been demonstrated in the MMI through several studies with regard to the attributes being assessed [9]. The domains identified through this study might explain why attributes for these subjective assessments have low reliability within stations, but moderate reliability across stations. For admissions committees, this means that the MMI is evaluating a broad set of skills rather than content-specific skills.

This study was limited to two US schools that use very similar, both with lengthy rater training orientations. It neither takes into consideration which scenarios produced which comments nor did this study attempt to examine the role that gender might play in the assessment or comment section. The scenario design is limited because both schools use ProfitHR MMIs. This study did not consider the rater background, number of years rating, or any other factors regarding the comments chosen. It is also possible that the training and explicit use of the comments during the process has influenced that actual comments. Perhaps a quantitative study at a school(s) that do not use the comments might yield different results. The overall score(s) were not considered in this study; thus, there are attributes that might be more associated with higher or lower scores. Further investigations are warranted.

Raters' MMI comments provide the context for numerical scores allowing admissions committees to more fully understand candidate strengths or weaknesses. Identifying the domains in MMI rater comments can ultimately assist the admissions committee to more comprehensively understand elements that raters consider as important to measure or identify during the MMI. These domains can also help to understand the reliability of multiple MMIs, even when some schools have tried to use MMIs to measure separate and different attributes in or across stations. Continued research at a qualitative level might be able to identify the various weight these domains contribute to MMI scores.

Future research could identify the weight rater comments have on the admissions committee decision. Also, it is not known if certain domains are related to higher or lower MMI scores, or if a relationship exists between the gender of the rater versus the gender of the prospective student.

Compliance with Ethical Standards

References

1. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. Med Teach. 2013;35(9): e1447–63.
2. Ponnamperuma G, et al. The long case and its modifications: a literature review, vol. 43; 2009. p. 936–41.
3. Regehr G, Freeman R, Hodges B, Russell L. Assessing the generalizability of OSCE measures across content domains. Acad Med. 1999;74(12):1320–2.
4. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ. 2004;38(3):314–26.
5. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Acad Med. 2004;79(10 Suppl):S40–2.
6. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ. 2009;43(8):767–75.
7. Kumar K, Roberts C, Rothnie I, du Fresne C, Walton M. Experiences of the multiple mini-interview: a qualitative analysis. Med Educ. 2009;43(4):360–7.
8. Alweis RL, Fitzpatrick C, Donato AA. Rater perceptions of bias using the multiple mini-interview format: a qualitative study. J Educ Train Stud. 2015;3(5):52–8.
9. Zaidi NB, Swoboda C, Wang LL, Manuel RS. Variance in attributes assessed by the multiple mini-interview. Med Teach. 2014;36(9): 794–8.
10. Eva KW, Macala C. Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. Med Educ. 2014;48(6):604–13.
11. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: rater performance and behavior when considering multiple competencies. Teach Learn Med. 2016;28(1):41–51.
12. Axelson RD, Kreiter CD. Rater and occasion impacts on the reliability of pre-admission assessments. Med Educ. 2009;43(12): 1198–202.
13. Cunnington JP, Neville AJ, Norman GR. The risks of thoroughness: reliability and validity of global ratings and checklists in an OSCE. Adv Health Sci Educ Theory Pract. 1996;1(3):227–33.
14. Manuel, R.S. MMI training sessions for rater. Use of Comments, sessions 2014;1-9 (slide 33). Retrieved from https://med.uc.edu/medicalstudentadmissions/interviewday/training. Accessed 9 July 2014.
15. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res. 2005;15(9):1277–88.
16. Glaser BG, Strauss A. The discovery of grounded theory: strategies for qualitative research. New York: Aldine De Gruyter; 1967. p. 271.
17. Wilkerson L, Rose M. Learning from thenarrative comments of standardized patients during an objective structured clinical examination of fourthyear medical students. ERIC Clearinghouse on Assessment and Evaluation. 2001.