Article:

Roberts, C. orcid.org/0000-0001-8613-682X, Burgess, A., Mossman, K. et al. (1 more author) (2025) Professional judgement: a social practice perspective on a multiple mini-interview for specialty training selection. BMC Medical Education, 25 (1). 18. ISSN 1472-6920

https://doi.org/10.1186/s12909-024-06535-3

Roberts *et al. BMC Medical Education* (2025) 25:18
https://doi.org/10.1186/s12909-024-06535-3

BMC Medical Education

RESEARCH **Open Access**

Professional judgement: a social practice perspective on a multiple mini-interview for specialty training selection

Chris Roberts[1*], Annette Burgess[2], Karyn Mossman[3] and Koshila Kumar[4,5]

Abstract

Background Interviewers' judgements play a critical role in competency-based assessments for selection such as the multiple-mini-interview (MMI). Much of the published research focuses on the psychometrics of selection and the impact of rater subjectivity. Within the context of selecting for entry into specialty postgraduate training, we used an interpretivist and socio-constructivist approach to explore how and why interviewers make judgments in high stakes selection settings whilst taking part in an MMI.

Methods We explored MMI interviewers' work processes through an institutional observational approach, based on the notion that interviewers' judgements are socially constructed and mediated by multiple factors. We gathered data through document analysis, and observations of interviewer training, candidate interactions with interviewers, and interviewer meetings. Interviews included informal encounters in a large selection centre. Data analysis balanced description and explicit interpretation of the meanings and functions of the interviewers' actions and behaviours.

Results Three themes were developed from the data showing how interviewers make professional judgements, specifically by; 'Balancing the interplay of rules and agency,' 'Participating in moderation and shared meaning making; and 'A culture of reflexivity and professional growth.' Interviewers balanced the following of institutional rules with making judgment choices based on personal expertise and knowledge. They engaged in dialogue, moderation, and shared meaning with fellow interviewers which enabled their consideration of multiple perspectives of the candidate's performance. Interviewers engaged in self-evaluation and reflection throughout, with professional learning and growth as primary care physicians and supervisors being an emergent outcome.

Conclusion This study offers insights into the judgment-making processes of interviewers in high-stakes MMI contexts, highlighting the balance between structured protocols and personal expertise within a socially constructed framework. By linking MMI practices to the broader work-based assessment literature, we contribute to advancing the design and implementation of more valid and fair selection tools for postgraduate training. Additionally, the study underscores the dual benefit of MMIs—not only as a selection tool but also as a platform for interviewers' professional growth. These insights offer practical implications for refining future MMI practices and improving the fairness of high-stakes selection processes.

*Correspondence:
Chris Roberts
chris.roberts@sheffield.ac.uk

Full list of author information is available at the end of the article

Keywords Selection, Postgraduate, Specialty training, Performance assessment, Professional judgment, Social learning, Professionalism, Work-based assessment, Multiple-mini-interview

Introduction

Interviewers' judgements play a critical role in competency-based assessments for selection such as the multiple-mini-interview (MMI). The making of professional judgments about the performance of those in professional training is attracting increasing theoretical and empirical attention in the assessment literature with a growing body of research [1–6]. Whilst selection into specialty training represents a high-stakes assessment context, much of the published work focuses on the psychometrics of selection and the impact of rater subjectivity. There are few studies that explore the professional judgement of interviewers in such contexts from a theoretical perspective. This study applies interpretive socio-constructivist theory to address this gap by exploring interviewers' judgements in high-stakes selection into general practice specialty training using a multiple mini-interview. As a point of language, General Practitioners (GPs) in specialty training in the UK, Netherlands and Australia are known as registrars rather than trainees.

Interviewer judgement in the multiple mini-interview

The multiple mini-interview (MMI), derived from the Objective Structured Clinical Examination (OSCE) has been extensively implemented in the undergraduate and graduate entry setting [7–9], and in postgraduate training selection in a range of international settings [10–17]. While many MMIs retain a structured format similar to OSCEs, in postgraduate contexts they are often tailored to reflect real-world clinical challenges, assessing competencies more closely related to clinical practice, for example offering insights into candidates' ability to navigate professional dilemmas [18]. Findings from a six-station MMI for postgraduates differentiate between two types of MMI questions: situational, which assess future behaviour in hypothetical scenarios, and past-behavioural, which focus on actual experiences. While situational questions reveal problem-solving abilities, past-behavioral questions like the approach used in our study, provide insights into candidates' real-world performance at work [17, 19]. One recognised issue with the fairness of MMIs in any setting is the degree of rater subjectivity [8, 9, 11, 14, 20–22]. To ensure the scores from observational assessments have validity, it is important to understand the underlying factors that influence raters when judging the abilities of candidates [2, 5, 23, 24]. Traditionally, the dominant approach for investigating rater influences has been through psychometric studies using a range of sophisticated regression techniques. Their findings are largely constrained by not being able to separate out rater issues from the station they are marking on [25]. Qualitative insights into this issue are rare but could better inform changes to the MMI design and interviewer training to enhance the quality of the selection process. Research into the variability of interviewers' judgement in MMIs suggests that for selecting into residency, interviewer self-perceived biases included cultural factors, personality factors, perception of prior preparation, concerns with norming, and biases associated with specific applicant characteristics [24]. This is not dissimilar to that found in the student selection setting, with similar issues for interviewers making their decisions about entry into a medical program [26]. These include conflict between independent decision-making and their need for a consensus around the expected standards for entry-level students, some uncertainty as to what they were assessing, recognising and addressing their subjectivity towards certain candidates, concerns over 'failing' candidates, and addressing candidates' use of impression management skills. Interviewer variability could be explained by their spontaneous application of subjective criteria (e.g., resilience) reflecting their taste for individual candidates [23]. Other researchers have found that assessors rely on global impressions informed by personal values, when distinguishing between similarly performing candidates, but did not explain why this might be so [27]. Moreover, recent studies on MMI rater cognition suggest that first impressions and other cognitive biases, far from being mere errors, may play a role in how raters form judgments, and significantly affect overall candidate evaluations [28]. This underscores the importance of understanding MMI interviewers as engaging in a process that blends intuitive judgments with structured assessment criteria.

Conceptual framework

Our conceptual framework outlines the key concepts, variables, relationships, and assumptions that guided our research study. We draw on literature from work-based assessment, principles of assessment for learning, and broader assessment practices to provide a foundation for the study's design and interpretation of results.

We came into this study, aware of the psychometric claims of objective high stakes selection, as might be typical of an OSCE style selection approach for students. That there is only one truthful reality in assessing any performance, and that measuring that reality results in a true score. Rater biases are considered as an error which can potentially be corrected through training [21, 29, 30].

Sensitised to the possibility that behavioural MMIs were about work practices, we took an interpretivist socio-constructivist approach to qualitatively explore explanations for interviewer behaviours within high stakes selection This philosophical and methodological approach emphasised the subjective nature of reality and the importance of social and cultural contexts in understanding our research phenomena. We viewed assessment as a social process [31] and professional judgement as a complex and interdependent process in moving from observation to judgement to rating to feedback [32]. Given the lack of theoretically informed literature on professional judgement in interviews, our initial framing prior to collecting the data was based on the work-based assessment literature [3, 4, 33–36], and the possibility of multiple "true" performance scores [33]. In this paradigm, assessors with different perspectives rate differently because they observe and value different aspects of performance or communicate information about the attributes measured in performance assessments [35, 37, 38] and their perceptions of the assessment task, and the context of the assessment [4]. Assessors are aided by but not reliant on rating scales in coming to their own independent expert judgment, particularly when making a decision on a complex performance [36, 39, 40]. In the broader assessment arena, there is a general scepticism that further training of judges makes much impact on reducing these systematic sources of error in their interviewing performance [26, 41], although some approaches for giving specific feedback back to assessors are thought to have promise [42, 43].

Study aims and research questions

This research addresses the gap in the literature by investigating professional judgement in high stakes contexts of interviewing for selection into specialty training. Specifically, we asked the research question. "How do interviewers make professional judgments in the context of high stakes selection into postgraduate specialty training within a multi-mini-interview approach?" This question is important because there is considerable resource and effort across a range of sectors in conducting MMIs. Answering this question will help selection and admissions designers optimise selection tool design and interviewer training in this context.

Methods

Drawing on socio-constructivist interpretations of MMI interviewer behaviour, this study employed a qualitative methodology. We used a method grounded in institutional observation [44–46] which examined the work based processes and studied how they are coordinated. Typically through texts and discourses of various sorts. In the context of MMIs, an institutional observational study aimed to understand the broader organizational context in which MMIs take place, exploring factors such as institutional policies, resources, and the overall functioning of the selection process. Thus, we addressed all the work done in the setting, including before, during and after the MMI interviewing, noting which activities were recognised and described institutionally, and which were not.

The setting and participants

The setting for this study was a single assessment centre within a national selection process, whose purpose in 2015 was to select eligible candidates into general practice specialty training using a combination of the situational judgement test (SJT) and the MMI. We have previously described aspects of this elsewhere [10, 16, 47]. This system was designed by external consultants. Typical of large-scale implementations, the MMI in this study consisted of four circuits per day (two simultaneous circuits at any one time). Each circuit consisted of the same six questions, with each question being asked by a single interviewer, resulting in each candidate having a total of six interviews with six different interviewers. Participants in this research were MMI interviewers and were largely full or part time general practitioners, and who were also GP training supervisors (including some medical educators), with a small number of non-clinical educators who had a management or educationalist background.

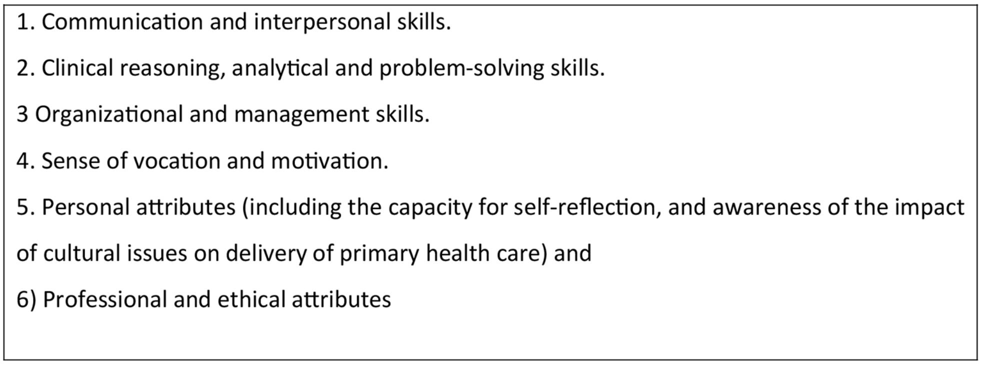
Ethics and consent

All research methods were conducted in accordance with relevant guidelines and regulations. The University of Sydney Human Research Ethics Committee approved the research. Protocol Number 13859. Informed written consent for participation was obtained from participants to enable us to include their data from this study. All participants were reassured that data were strictly de-identified to protect participant privacy.

The MMI process

Interviewer training

All interviewers were offered face to face training and support materials before the interviews. The focus was on behavioural interviewing techniques to elicit examples of times when the candidates demonstrated the required behaviours, using prompting and probing questions. Probing questions were of the "How," "What," "Why," and "When" variety. Interviewers were strongly discouraged from using "What if" questions, since these raised a new hypothetical situation. Before the interview, the lead medical educator briefed each set of interviewers about the MMI with a focus on the rating scale and calibration, making notes, avoiding bias, fairness, the logistics of the



1. Communication and interpersonal skills.

2. Clinical reasoning, analytical and problem-solving skills.

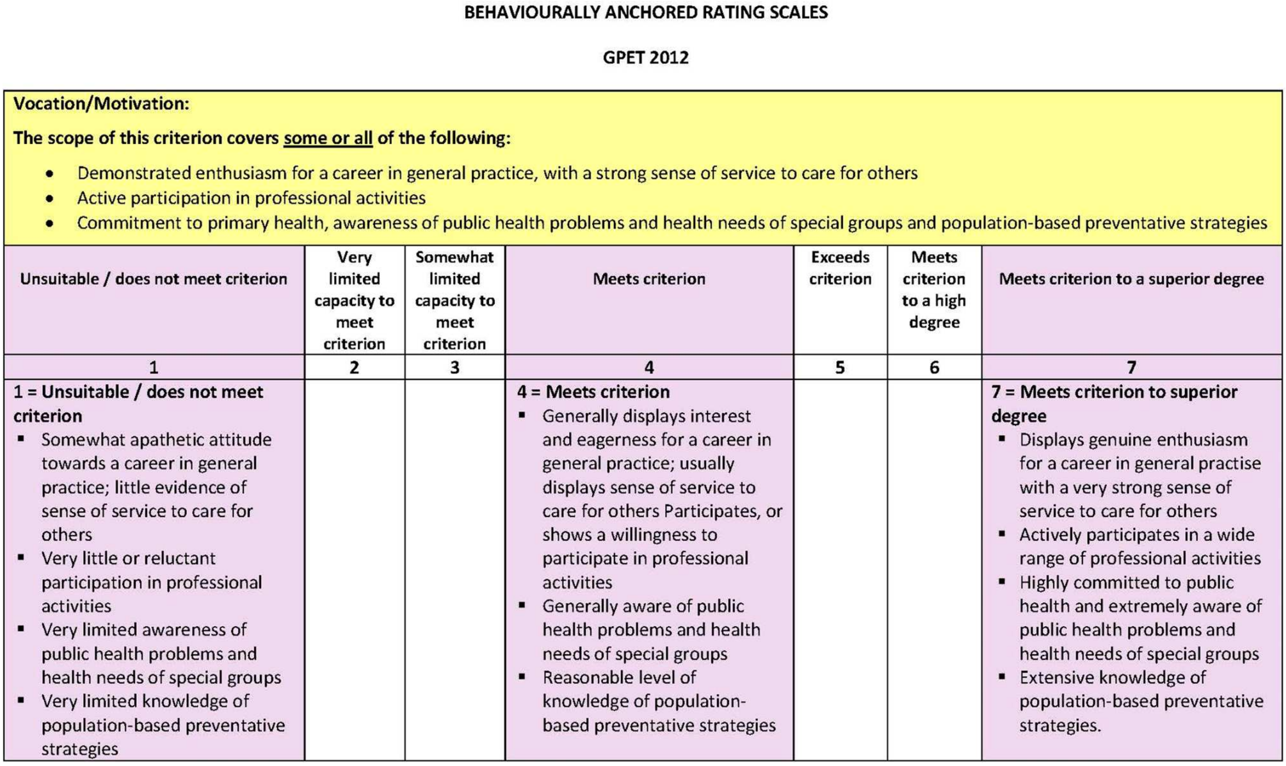
3 Organizational and management skills.

4. Sense of vocation and motivation.

5. Personal attributes (including the capacity for self-reflection, and awareness of the impact of cultural issues on delivery of primary health care) and

6) Professional and ethical attributes

Fig. 1 Domains of practice that underpinned the MMI in the research context



BEHAVIOURALLY ANCHORED RATING SCALES

GPET 2012

Vocation/Motivation:

The scope of this criterion covers <u>some or all</u> of the following:

- Demonstrated enthusiasm for a career in general practice, with a strong sense of service to care for others
- Active participation in professional activities
- Commitment to primary health, awareness of public health problems and health needs of special groups and population-based preventative strategies

Unsuitable / does not meet criterion	Very limited capacity to meet criterion	Somewhat limited capacity to meet criterion	Meets criterion	Exceeds criterion	Meets criterion to a high degree	Meets criterion to a superior degree
1	2	3	4	5	6	7
1 = Unsuitable / does not meet criterion ▪ Somewhat apathetic attitude towards a career in general practice; little evidence of sense of service to care for others ▪ Very little or reluctant participation in professional activities ▪ Very limited awareness of public health problems and health needs of special groups ▪ Very limited knowledge of population-based preventative strategies			**4 = Meets criterion** ▪ Generally displays interest and eagerness for a career in general practice; usually displays sense of service to care for others Participates, or shows a willingness to participate in professional activities ▪ Generally aware of public health problems and health needs of special groups ▪ Reasonable level of knowledge of population-based preventative strategies			**7 = Meets criterion to superior degree** ▪ Displays genuine enthusiasm for a career in general practise with a very strong sense of service to care for others ▪ Actively participates in a wide range of professional activities ▪ Highly committed to public health and extremely aware of public health problems and health needs of special groups ▪ Extensive knowledge of population-based preventative strategies.

Fig. 2 Example of behaviourally rated anchor scale (BARS) for a single MMI question about a sense of vocation and motivation in becoming a GP

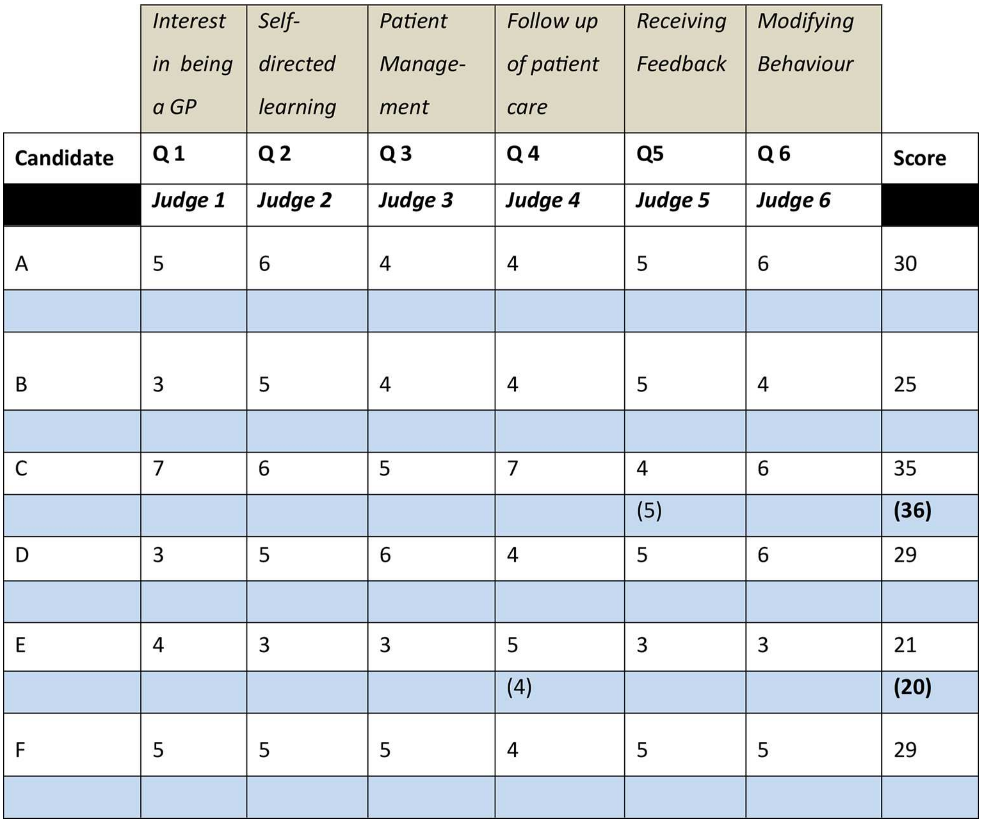
circuit including interviewers' post interview discussion meetings, confidentiality and refreshments,

The MMI rating scale

Selection criteria for the range of expected skills and behaviours at entry into general practice training were based on six competency domains (see Fig. 1). The assessment scale consisted of a behaviourally anchored rating scale (BARS), which was customised for each domain. For example, the vocation and motivation question illustrated in Fig. 2, included a descriptor of the criterion e.g., enthusiasm for a career in general practice, which was marked using a seven-point rating scale. This scale ranged from 1 (unsuitable/does not meet criterion) to 4 (meets criterion) through to 7 (meets criterion to a superior degree). For each anchor, descriptors were provided to illustrate the different ways in which candidates might meet the criteria in the interview (see Fig. 2).

During the interview

Candidates had two minutes to reflect and make notes on the question before entering the interview room. Prompts were read out as written in the MMI question or paraphrased, and sometimes repeated for the applicant. Interviewers posed questions, with varying degrees of skill, using the behavioural 'probes.' For example, interviewers responded to a candidate answer by asking "tell me more," or "what did you do in that situation?" to elicit



	Interest in being a GP	*Self-directed learning*	*Patient Manage-ment*	*Follow up of patient care*	*Receiving Feedback*	*Modifying Behaviour*	
Candidate	**Q 1**	**Q 2**	**Q 3**	**Q 4**	**Q5**	**Q 6**	**Score**
	Judge 1	***Judge 2***	***Judge 3***	***Judge 4***	***Judge 5***	***Judge 6***	
A	5	6	4	4	5	6	30
B	3	5	4	4	5	4	25
C	7	6	5	7	4	6	35
					(5)		**(36)**
D	3	5	6	4	5	6	29
E	4	3	3	5	3	3	21
				(4)			**(20)**
F	5	5	5	4	5	5	29

Fig. 3 De-identified marking matrix for single circuit of MMI questions by interviewer, candidate and MMI question type

the evidence they were looking for to apply the rating scale. At the end of the interview, the interviewer completed the rating scale as a number from 1 to 7 and made notes justifying their decision.

Post interview scoring

The MMI system had a unique component compared with other selection settings. At the end of each MMI circuit, interviewers met to discuss their marks facilitated by a senior medical educator. The aim of these sessions, for which prior evaluations had reported widespread support, was educational. it gave interviewers the opportunity to gain greater insights into the nuances of applying the rating scale. In the discussion, a matrix was presented by administration staff which contained all the applicant scores for that circuit in a 6 by 6 table with individual interviewers as columns and the candidates as rows. An example of this is given in Fig. 3. In the interests of time, discussions centred around scores of less than four (i.e., limited capacity to meet the criterion) or where there were three marks difference between the lowest and the highest score given for a candidate. Interviewers shared challenges raised by particular candidates, and discussed, and defended or moderated their decisions. After the discussion, interviewers could indicate whether they would have changed their score, though the original score prior to the discussion was used. Interviewer meetings varied from 15 min to half an hour depending on the issues of concern.

Data collection

We collected data via observations in a single assessment centre which operated over two days, where 43 interviewers interviewed 243 candidates. Using national survey data, of the 2,154 candidates who took the MMI in the year of the study, females (59%) outnumbered males (41%). Their age ranged from 22 to 68 years, with a median age of 30 years. Seventy-one per cent of candidates were an Australian Medical Graduate (AMG) or a Foreign Graduate of an Accredited Medical School (FGAMS). Overseas Trained Doctors (OTD) made up 29% of the candidate pool. Nearly half (42%) of candidates did not have English as their first language. Of the interviewers nationally, a third of interviewers were aged 50–59 years. Females were 56% and the majority (86%) were from an English-speaking background.Forty percent of interviewers had their primary medical qualification from Australia or New Zealand. 6% from the UK and 3% from India.

Our observation fieldwork captured various sorts of behaviours from participants through the prolonged involvement [48–50] of three authors (CR, AB, and

KM). This included field notes from observing over two days, four pre-interview briefings for interviewers and five for candidates, twelve post-interview meetings for interviewers, and 60 MMI candidate interviews. Data also included interviewer training manuals, participant orientation protocols, and access to completed rating forms of individual examiners in the sessions. In the pre-interview briefings, we observed how guidelines shaped fairness and consistency expectations. During the MMI, we recorded both verbal and non-verbal interactions, focusing on the extent to which interviewers followed the provided instructions, noting where, how and why they deviated from the guidelines in coming to their judgements. Post-interview debriefings provided insights into how interviewers synthesized station results and reached consensus on rankings. Field notes were taken by hand in a semi-structured format to capture the nuances of interviewer behaviour and decision-making. Daily reflective discussions allowed the team to analyse observations and resolve interpretation differences, ensuring a comprehensive understanding of the MMI process and professional judgments.

Opportunistically, we interviewed 19 interviewers over two days, using a semi structured interview guide based on assessment principles from the literature including, training, motivation, fairness, acceptability, transparency, alignment with marking criteria, potential biases, decision making, and feedback. Of the 19 interviewed, 12 were male and 7 were female, 12 were overseas trained doctors with 7 being Australian or New Zealand trained. All interviewers were fluent in English. Additional interview questions were added as interesting patterns in the data emerged. Interviews lasted a mean of 15 min (range 4–28 min), and all interviews were audiotaped and transcribed verbatim. Reflection on the structured interviews, as well as observation of interviewers who were not interviewed were included in our notebooks.

Data analysis

During the observation phase, prior to formal analyses, we conceptualised the assessment centre as a place of work for interviewers and a space for candidates to show job related skills. A constructivist–interpretivist approaches sensitised by the work-based learning and assessment literature seemed appropriate to capture and understand contextualised learning and performance in work settings [3]. Data analysis involved explicit interpretation of the meanings and functions of human actions, the product of which mainly took the form of verbal descriptions and explanations. The initial analysis primarily focused on exploring the social relations that people were drawn into through their work [44–46]. Data were displayed so as to balance the requirements of description and analysis, and encourage researcher reflexivity [51]. Data were triangulated from field notes e.g., relating expected behaviours from document analysis to direct observations of interviewer-candidate encounters, including comments from interviewers on their thoughts and feelings of that MMI station. Further triangulation matched field notes of the interactions within the interviewer meetings with the visual display of candidate assessment scores. We constructed a rich description of how interviewers made their judgements and the decision-making process they engaged in, by moving back and forth between the raw data and the theoretical understandings drawn from the work-based assessment literature. Consistent with the principles of inductive analysis, we developed three themes relevant to our specific research questions [52] after negotiation between the authors.

Team reflexivity

Using Barry et al., [53] as a guide, our research team regularly reflected on how aspects of ourselves, our teamwork, and our methodology impacted our interpretations of participants' experiences. Three of the team (CR, AB, and KM) had worked severally as a team conducting evaluations of the MMI in other centres around the country. CR, an academic general practitioner, former GP trainer, and medical education researcher, provided valuable insights into the clinical context and what it means being a GP and a trainer [54]. Both AB and KK had nonclinical educational research backgrounds. KK joined the research study bringing additional expertise in qualitative research, adding depth to our insights into the complexities of professional judgement. KM was a senior administrator and evaluator with extensive experience of managing the logistics of large-scale assessments. Following the selection centre days, reflexivity was enhanced through ongoing meetings and email discussions, where we shared deliberations and resolved conflicts in data interpretation.

Findings

We developed three major themes in relation to our research question which were 'Balancing the interplay of rules and agency,' 'Participating in moderation and shared meaning making,' and 'Culture of reflexivity and professional growth.' We summarise each theme and illustrate our description and interpretation with the voices of the interviewers.

Balancing the interplay of rules and agency

This theme describes how interviewers balanced on the one hand their role as an objective interviewer who follows a set of institutional protocols and rules, and on the other hand their own agency and professional insight, knowledge, and experience to make decisions.

In the MMI interview with candidates, interviewers appeared comfortable following the rules of the behavioural interviewing style but were surprised how emotional and subjective the interviewing could be, and how those subjective emotions impacted their objective judgement. Interviewers often found that candidate stories in response to an MMI question such as "tell me a time when…." were emotionally engaging and at times upsetting. Interviewers gained a privileged insight into the world of early career doctors and the significant challenges of working in the public health system. They were more able to make sense of a candidate's responses relative to the context in which the story was told. For example, in response to an MMI question about times the candidates were 'impacted by your own health" common stories from applicants included difficulties in clinical decision making, being impacted by long shifts, unexpected cover for sick colleagues, working following serious personal difficulties, and being subjected to bullying and harassment.

Most interviewers appeared to hold a positivist view of candidate performance, believing that they had provided a single 'true' performance for each candidate, unimpacted by error or bias, for example their own subjectivity, context specificity or question difficulty. Interviewers felt the writing of detailed notes was essential to justify their judgement, but a variety of note taking styles were used, ranging from focusing on borderline candidates, to recording everything verbatim sometimes to the exclusion of eye contact with the candidate. They were aware of the need to avoid the common interviewing pitfalls, which had been outlined in their training, such as comparing candidates with each other or themselves. Interviewers described the tensions between being subjective and objective and hinted at the methods of adjustment they used to take account of their first impressions of a candidate to avoid miscategorisation within the rating scale.

> *I know that if I see a candidate. …that reminds me of me when I was that age, but the BARS (rating scale) keeps you on track.*

For less strong candidates, many interviewers wanted to record their specific thoughts about the likely professional support needed for the candidate progression in future training, a form of elaboration of their judgement. This was particularly where the interviewer thought a candidate borderline in meeting the selection criteria.

> *Most applicants give mixed information, we (supervisors) need to know that what is in between would cause the most angst. Interviewers need to understand applicants don't fit exactly. The judge needs to make a balanced judgement based on the quality of information they gave regardless of nervousness or garrulousness. This relates to the importance of notes.*

Despite general adherence to interviewing by the institutional rules, some interviewers chose to go beyond the rules to get responses to inform their rating of a candidate. This often occurred where candidates failed to authentically engage with the MMI question, despite prompting and probing. For example, to rate the MMI question around the "values and commitment required by the candidate" interviewers were judging candidates' "enthusiasm for general practice." Interviewers went beyond the rules to separate issues of candidate impression management or lack thereof and a genuinely authentic answer.

> *One person for instance spoke in a bit of a monotone and wasn't particularly charismatic or dynamic. But through the answer he was really, really clearly enthusiastic about GP as a profession - but he didn't present himself enthusiastically. So, it was - it was interesting, sort of, picking the difference between that and somebody who presented really enthusiastically, but maybe who hadn't put as much thought into what about general practice appeals.*

Interviewers reflected on factors for which they regularly broke the rules and adjusted for when making their own judgement. For example, by rating for matters that were generic and not related to the MMI question. This included candidate nervousness, self-confidence, talking a lot and repeating themselves, appearing overcoached, the degree to which candidates had to be prompted, and English not being a first language for some candidates.

Experienced interviewers often talked about having a gut feeling about candidates, which didn't require them to analyse the performance according to the structure given by the rating scale. One quipped "*We use the gut…. better than the BARs*" (rating scale). They were reflecting on candidate performance from their perspective as an experienced supervisor or a medical educator. For example, one interviewer described the suitability of a candidate to be a GP.

> *The candidate was speaking very clear…. her volume was so good that it didn't – didn't look that she is shouting at me, or she is mumbling. So that gave me a gut feeling in this – that this person, if she's thinking to be a – a general practitioner, she will be one.*

Sometimes interviewers who went beyond the rules because they felt the MMI question needed to change to better capture what they were looking for in a candidate. This most often related to misalignment of the selection criteria with the MMI questions and the written prompts. Some interviewers noted the use of probes in the behavioural interviewing style was somewhat like a consulting style with their patients and similarly took experience to master.

In summary, for many interviewers there was a delicate balance between interviewing by the rules and needing to go beyond the rules. Interviewers appeared engaged in the behavioural interviewing process and used it with varying degrees of skill. New interviewers were rapidly socialised into the format and expectations of the MMI. There was a strong affective element that informed interviewers judgment through the stories of practice related by the candidates answering the MMI questions. In the face of uncertainty, misaligned questions and personal intuition, the more experienced interviewers often broke the rules to get the evidence they needed to inform their judgements of candidates using their professional insights and knowledge. Some interviewers broke the rules because they did not fully understand the importance of staying within the question andmaking consistent judgements.

Participating in moderation and shared meaning making

This theme describes the act of communicating and justifying, defending, or adjusting their decision about a candidate's suitability for training through a dialogic process. It encompasses the sharing of the combined and total marks for the candidates as an aid to calibration.

Following the completion of each circuit of the MMI, interviewers reviewed the matrix showing the individual marks, by interviewer and by candidate (Fig. 3). Interviewers spoke to their most compelling evidence to support their judgement particularly if it appeared to be different from that of others. While there was no collective decision making by interviewers, there was an opportunity to discuss, moderate, and refine their own individual judgements, and affirm an overall consensus that the score and its associated qualitative comments were fair. By the time that interviewers came to the meeting they had already decided on a candidate score. In the discussion, they most often referred to the numerical score rather than qualitatively describing the extent to which the candidate had or had not met the criteria. They consistently re-iterated that the candidates they had seen were a "f*our*" or a "*five*", i.e., meets criterion or exceeds criterion and then justified their score using explanatory comments. For example, one interviewer suggested about a candidate.

> *She had done enough to be a five, but she couldn't give me any more to be a six.*

The question arose as to what extent the interviewers were too quick in their reasoning about a candidate's ability and would put them 'into pre-existing schemas' [35], an issue that has been identified for assessors in work-based assessment. If an interviewer acknowledged miscategorisation might have happened, there was an opportunity in the post interview discussion meeting for immediate reflection on the common reasons why. This was more likely to happen with novice interviewers or in the first circuit of a session, where an interviewer was developing an understanding of a particular MMI question. The meeting moderator facilitated such interactions, and interviewers might respond to a marking decision of another with comments such as, "*you were a bit harsh?"* suggesting some honest disagreement. More commonly, the facilitator reassured the interviewer that they had given a good defence of their decision. The interviewers were allowed to include their adjusted score (see example of score in brackets in Fig. 3) but it was their original score that counted. In the discussion, where an interviewer thought another was biased, they tried to correct the speaker's misconceptions of scoring by referring them back to the BARS and the training materials. Reflecting on the possible bias of others allowed interviewers to reflect on their own biases.

> *In our debriefing you heard people saying, 'well, this person was really pleasant.' So, it's having to remind myself not to score them, you know, a six or a seven based on the pleasantness and to make that maybe a five instead of a six, or a six instead of a seven, um, when it's correcting for enthusiasm or pleasantness, or something else like that. So, I think other people are mindful of it, but it's something that if you're not, I think makes the mark creep up. …Or creep down.*

In the discussion meeting, interviewers received feedback and heard about what other interviewers had thought about the candidates. This reassured them about the consistency and fairness of their own judgment. One interviewer was alerted by apparent over confidence in a candidate but was reassured by triangulating the candidate's body language, her talk, and later by hearing what other interviewers had decided.

> *…. she (the candidate) was …. very confident. And her body language was different according to what she was telling. Which meant she was – she was truthful with things. She was …able to pass a smile. …And that gave me a gut feeling that means she is not just making it up…. And the good thing was that when we, ah, looked at the score (in the debriefing meeting) she was on the top in every part.*

The dialogic aspect of the interviewer meetings offered a way of collective sensemaking about challenging decisions. Interviewers shared and discussed how they were responding to recurring and major behavioural, or knowledge issues raised by the interview that were unrelated to the set marking criteria. Examples included where the candidate demonstrated a lack of insight into their own deficient problem-solving skills or demonstrated potentially challenging behaviours such as arrogance, insufficient competency, lack of insight into the impact of strongly held cultural beliefs on patient care, and global language issues. One commonly agreed work around was to encourage the interviewer to agree a form of words to add to the comment sections of the spreadsheet as a way of registering concerns about candidates. For example, an interviewer's sense of "arrogance" in the candidate might be noted as "someone who may be challenged in accepting feedback."

In the meeting, interviewers took care to demonstrate awareness of their own potential biases on matters that would be highly inappropriate to scale. For example, a female candidate who wanted to defer to her husband on a matter of professional judgement, or a male candidate who wished to bring patients into a closer relationship with his god. The challenge of decision-making was heightened for candidates whose first language was not English, particularly for overseas-trained doctors. Interviewers grappled with balancing the potential impact of communication on future patient care while ensuring they did not discriminate based on language or cultural differences.

> *People's ability to communicate clearly in English, um, and so, it, so there's always the – have they understood the question, um, or are they just having difficultly because English isn't their first language? So, we had a few of those where people sort of said – they gave an answer but, um, you know, a lot of syntax errors. Um, but if that's in a communication station, that's perfectly legitimate.*

Some interviewers were not doctors and came from a managerial or educational background and thus offered a different professional context, background, and experience. They differed from the doctors in the interviewer meeting in how they explained their professional judgement and their justification of it. Doctors appeared to frame much of their impression of the candidate in terms of how they would behave in the workplace. This was sometimes discussed in meetings as being practice ready, although the stated purpose of the selection process, including the MMI, was to determine the trainability of candidates. Administrators talked more to their knowledge of how GP registrars in their area behaved, professionally. Collective sensemaking during the discussion meeting could have negative impacts. Sometimes medical educators were

> *Put off by the group discussion as well as all the biases that come out and the disclosure that people had gone off track. People who do it (the interviewing) a million times are set in their ways and difficult to mould and manage.*

However, the wider group of interviewers acknowledged that the differing interviewer backgrounds brought a different and valued perspective to the collective understanding of a candidate's capabilities.

Overall, interviewers had a shared understanding of applying the numerical scoring system and used the interview scale categories to evaluate candidate suitability for training. Qualitative comments were used to justify the interviewer's score. Discussing the matrix of candidate scores for that round, including the total candidate score, assisted the certainty of the decision-making process. However, it could be misleading in implying that the numbers reflected a true score rather than a biased one. Some interviews felt they had miscategorised candidates, and wanted their adjustment recorded. Discussion with their peers about the rationale of their decisions enabled interviewers to improve future decision-making. There were a large range of issues that were important to interviewers in their individual scoring, but not included in the rating scale. Here, collective sensemaking and moderation of decisions was based on sharing personal experience.

A culture of reflexivity and professional growth

This theme described the self-evaluation and reflexivity of interviewers throughout the MMI process about their role in learning and teaching and assessment, as interviewers, as supervisors, and as clinicians. It includes some emergent outcomes from the overall MMI participation experience.

Most participants found the dialogic discussion with their peers allowed reflection on both their own performance as an interviewer and that of others in the context of dilemmas in professional judgement. For novice interviewers, professional expertise development was facilitated by reflecting with more experienced colleagues to enrich their understanding of the nuances of the selection process. The inexperienced interviewers demonstrated the largest growth curves, including their anxieties about being a first-time interviewer.

> *It gives confidence that you are, when you listen to what the others say, you are sticking to the criteria.*

> *It also lets you know if you are being too strict or too soft.*

Discussions with colleagues over two days about candidate performance and suitability for specialty training facilitated shared reflection, sensemaking and learning related to the experience of supervising the registrar whose performances was causing concern. To try and prevent such registrars from entering postgraduate training was a motivation for many to be involved in the selection process. Though they recognised that some of the problems with the registrars were difficult to pick up with any selection tool or combination of them.

> *My second example was female, mid-thirties, overseas doctor. Her biggest hurdle was lack of insight into her deficiency. No analysis, no critical thinking in a differential diagnosis. She failed (professional training college knowledge test) but had passed the MMI. How do you pick up trainability, self-reflection, and insight?*

Interviewers were at different points in their career both professionally and as an interviewer with varying degrees of insight into the interview process. Many interviewers had experience of assessing in other situations for example with their professional college or with the university and were "*comfortable in the space*" of standardised approaches. More experienced interviewers experienced the discussion meeting as one of professional growth in a different way. They gained much from validating their performance as an interviewer, but also from reflecting on problem decisions. For example, where candidates expressed stories of compromises of patient safety or of working despite personal health issues. Interviewers found this process valuable for their own continuing professional development as a general practitioner as well as a GP supervisor.

> *actually doing it in practice, I think you can only really understand that by actually doing it on the day, because, you know, even, sort of – I think we went through, maybe, three or four, sort of, hypotheticals, um, but again, you don't, sort of, really appreciate the diversity and how it all unfolds until the actual day...because, you know, there's so much variability in, you know, the responses from the registrars, I – I don't think any training will actually fully equip you for that".*

An important factor that interviewers talked about as well as their own growth as interviewers and supervisors, was the desire to assess the future growth of candidate. The rules of the MMI emphasised the collection of information to assess a candidate's attainment thus far. However, interviewers discussed predicting what that candidate might be like in future situations in general practice, a hypothetical "what if," which the selection rules discouraged. For example, in dealing with a candidate who was not yet ready for the training program.

> *[The candidate] with the right supervisor I think will grow into this and I think be a good registrar, it's just they are not quite there yet.*

In summary, learning about and developing expertise in making professional judgment in the context of selection was a social activity conducted in interaction and participation with others. This was important for learning about aspects such as the core values of general practice, supervising the registrar whose performances was causing concern, and the complexities associated with making high stakes professional judgements. Having from 4 h to two days together provided opportunities to reflect on what it means to be a GP supervisor, and what it meant to be a primary care clinician. This notion of a learning culture was an emergent outcome of the selection process, i.e. learning that was greater for interviewers overall than from the individual components of learning.

Discussion

Summary of key findings

This study provides insight into how interviewers make professional judgments during high-stakes selection in MMIs for postgraduate specialty training. Interviewers balanced institutional guidelines with their professional judgment, frequently deviating from rigid protocols to better understand candidates, particularly in emotionally charged situations. This highlights the limitations of structured frameworks in fully capturing the nuances of real-world clinical experiences, which is central to both socio-constructivist and work-based assessment theories.

A key finding is that interviewers blended subjective impressions with formal assessment criteria, particularly when assessing borderline candidates. This shows that judgments were shaped by personal experiences and context, echoing work-based assessment principles, where professional judgment is fluid and responsive to situational factors rather than strictly objective. The study also revealed that dialogic moderation during post-interview debriefings was important. These discussions allowed interviewers to engage in collective reflection, reducing biases and leading to more consistent, well-rounded evaluations, akin to reflective practice in workplace assessments.

Additionally, interviewers' reflexivity contributed to their professional growth, as they became more aware of their biases and judgment processes through peer discussions—an important element in work-based learning, where continuous professional development is achieved through reflection and feedback.

From a socio-constructivist perspective, these findings emphasize that professional judgment in MMIs is socially constructed, shaped by the interaction between institutional guidelines, personal insights, and collaborative reflection. This challenges the assumption of objectivity in selection processes, suggesting that flexibility, reflection, and dialogic practices—core to both socio-constructivist and work-based assessment theories—are important for improving fairness and accuracy in candidate evaluations.

Comparison with existing theory and literature

Our findings extend the professional judgement literature in selection by linking the theory and practice of interviewing for selection with the relatively rich theoretical understandings in work-based assessment [55, 56]. Our findings reinforce the concept that past-behavioural MMIs align with key principles of workplace-based assessment (WBA). While OSCEs are known for their structured assessment of technical skills and WBAs for assessing real-world performance, past-behavioural MMIs blend these approaches. This hybrid model presents a challenge for interviewers who must balance standardized scoring with professional judgment in clinical practice [17–19]. Although direct comparisons between MMIs and WBAs are limited, our study suggests that MMIs, especially in high-stakes contexts, combine elements of both, contributing to discussions on rater-based judgments in postgraduate selection.

Our study suggests that acquiring expertise in the professional judgment of candidate behaviours in this context is social and reflexive in nature. This challenges the prevailing view that MMI interviews are objective as in an OSCE, and that raters are a source of modifiable bias. Interviewers' behaviours in the context of high stakes selection align with many of the characteristics of raters judging a complex performance in work-based assessment, using information related to a range of situational factors and personal experiences. The evidence-based approaches which have so far been discussed in the context of work-based assessment [2, 3, 6, 41, 57, 58] appear to provide a helpful framework to inform and enhance understandings of interviewers' decision-making in future iterations of the MMI.

Our findings on how interviewers balanced structured protocols with personal intuition align with emerging literature on rater cognition in MMIs. As seen in recent studies on the impact of first impressions [28], interviewers often made early judgments that shaped subsequent ratings. However, these initial impressions were moderated through post-interview discussions and reflective practice, leading to more holistic evaluations. This mirrors broader findings in performance-based assessments, such as Wood et al., [59], where raters adjusted their judgments based on evolving performance, blending structured frameworks with real-time adaptation.

Our findings suggest that, as in work-based assessment, interviewers in high stakes selection may use differing processes in judging candidate performance, in a way that has similarities to clinical reasoning i.e. making some instant and intuitive decisions about candidates based on pattern recognition [2]. This is an example of the fast or System 1 processing found in work-based assessments, where several factors such as presence and communication influenced interviewers' first impressions of candidates. In contrast, the rating scale in this MMI was designed to promote slow or System 2 thinking, which is analytic. Rather than being an unstable cause affecting.

individual performance as in work based assessment, the display of emotions by candidates provided authenticity and aroused empathy from interviewers [60].

Our study also highlights the importance of experience in interviewer judgment, with novice interviewers tending to be more analytical until they gain experience, while experienced interviewers were more likely to make decisions on gut feelings and then moderate this with further probing. Interviewers with different perspectives of specialty training for example nonclinical interviewers were comfortable to rate differently because they were observing different aspects of performance in their interview station. This would concur with the notion of assessing multiple 'true' performances [33].

Interviewers tended to rate candidates based on a global score derived from pre-set criteria, which may not necessarily reflect the full range of skills or attributes required for the job. This contrasts with the findings of Yeates et al. [58] who discovered that judgments in work-based assessments were mostly expressed using descriptive language. The interviewers in our study initially focused on numerical scores, which suggests that the use of numbers to assess a candidate's suitability for training had become socialised in the group's language and practices. Interviewers highlighted their most compelling evidence to support their decision, which is consistent with the findings of a study on raters in work-based assessments [54].

Our research also highlights the importance of developing a culture of collective self-reflection and self-evaluation for interviewers, as it can help them become more aware of their own professional growth as interviewers and supervisors and improve their judgement of candidates. This work extends the notion that such practical wisdom, largely acquired through experience and informal conversations with respected peers [61], can be supported through facilitated discussion meetings.

Methodological strengths and areas of uncertainty

Theoretically informed studies addressing how interviewers make professional judgments in the context of selection into specialty training. A crucial tension in observational studies in educational settings, is balancing the importance of description versus interpretation. That is understanding the perspectives of the people being studied and developing an analytic understanding of their perceptions, activities and action [62]. While the observational approach provided rich insights into real-world professional judgments within MMIs, it was not a formal ethnographic study. The episodic nature of the observations may limit replicability, though the consistency of field notes and reflections mitigated this concern. The methodology was well-suited to capturing the complexities of postgraduate MMIs, where interviewers assess professional work-related experiences rather than standardized OSCE-style interactions. By collecting the perspectives and work experiences of differing interviewers over two days we claim a sample with sufficient information power given the focused aims of our study [63]. The methodology effectively captured the complexities of postgraduate MMIs, focusing on professional experiences over standardized OSCEs. Although institutional norms influenced judgments, GPs typically operate in less hierarchical settings, making them less affected by top-down processes. However, the structured selection process in this study reflects the standardization necessary in high-stakes environments, enhancing the transferability and relevance of the findings to similar contexts [64].

A further consideration in observational studies is the pragmatic issues of balancing on the one hand, the disruption of collecting demographic details of all participants in time pressured settings and on the other hand demonstrating inclusivity of the sample. We acknowledge that the examiner meeting in this study is an unusual feature compared with most MII studies. We also acknowledge that we could have been more sensitive in demonstrating inclusivity and diversity by collecting more detail of the characteristics of the interviewers. However, with these caveats, the lessons learnt about the social aspects of assessment is adaptable to other settings. Finally, we acknowledge the long gap in the publication process, during which the manuscript was reviewed and later set aside. However, we revisited it due to the continued relevance of the findings despite the age of the data. The insights into the MMI process and interviewer decision-making still reflect key challenges in postgraduate medical selection, underscoring the lasting importance of this research for contemporary practice.

Implications for educational practice

There are several implications of this research in shaping the development of more effective training and support for MMI interviewers. This includes providing opportunities for reflection, discussion, and sense-making activities to enhance their judgment-making expertise. If admissions designers can better understand how MMI interviewers are making decisions both individually and collectively, they can optimise the rating scale to enhance decision-making processes.

First, despite the strengths of the behavioural MMI format, there are inherent limitations in its ability to assess certain key qualities, particularly future trainability. While MMIs are effective in evaluating candidates' current communication and decision-making skills, they may not fully capture a candidate's capacity for long-term growth or adaptability in clinical practice. Research on professionalism highlights the difficulty in predicting future professional behaviours and development through short, structured interviews alone [55, 65]. As a result, additional assessment methods, such as longitudinal assessments including workplace-based asessments, may be necessary to better gauge candidates' potential for growth and development over time.

Second, regarding faculty development, selection system designers could use work-based learning approaches to promote interviewers' professional judgement expertise. The form of work-based learning is very different from more traditional training approaches of providing interviewers with detailed didactic information or feedback related to their numerical scoring [41, 42, 66]. As in work-based assessment, it seems unlikely that current strategies of increased interviewer training underpinned by psychometric theory such as tightening marking criteria and re-enforcing awareness of bias would address interviewer subjectivity. More appropriate 'train-the-trainers' sessions could be renamed as 'assessor readiness' sessions provide continuing professional development underpinned by a constructivist learning and teaching models [41]. For example, interviewer examples about decision making dilemmas or complexities can be captured and integrated into the training, so as to orient interviewers to the realities of balancing objectivity and subjectivity. In a system not dissimilar to that used in video based OSCE examiner training, interviewers could assess videoed MMI performances, receive expert-panel ratings and justifications for a selection of the videos the interviewers had assessed and then discuss these with fellow interviewers [67].

Fourth, as in work-based assessment literature, improved decision making of interviewers will follow, if there is better constructive alignment [36] of the rating scale. One way of achieving this alignment would be to capture meaningful criteria from interviewers when developing new or revising existing MMI questions. This includes investigating ways of supporting the identification of candidates' capability for professional growth. We suggest that the focus at the postgraduate level in selection should be on determining potential for 'growth' in addition to "excellence in attainment to date." Finding a mechanism to determine a candidates' capability for growth could use the method of Crossley and Jolly [57] by determining how interviewers think this could be measured.

Finally, some of the issues that interviewers identified as barriers to entry into specialist in training were global communication difficulties, arrogance, lack of conscientiousness, problems with resilience, cultural safety and awareness, and unseen health problems. These are characteristics that some researchers have suggested are also an issue in medical school selection [68]. We recommend that any judgment-based instrument include a non-scaled narrative style question about concerns that should be noted by others, because there's frequently no way to reflect such concerns with any categorical rating scale [57].

Implications for further research

Interviewers are highly expert and often constrained in making good judgments by the limitations of the tools that they have at their disposal. More research on the impact of using methods developed through socio-constructivist understandings developed in work-based learning and assessment in the enhancing the professional judgement of interviewers is required. The methodology of institutional observational studies may be useful to researchers treating large scale assessment and selection activities as work processes and looking beyond basing findings on formal interviews by allowing consideration of texts and discourses of various sorts [45]. This methodological approach could be applied to other selection processes beyond MMI interviews and inform the development of more effective assessment practices that take account of the social and reflexive nature of professional judgement. Institutional observation studies could be useful in researching newer notions of assessment for learning in the postgraduate sector where subjective decision making is privileged [69–71]. Calls for an ethnographic approach to large scale assessment might suit researchers seeking to understand the entirety of a social setting rather than isolated variables enabling them to consider the interconnectedness of various aspects of culture, social structures, materials, and individual behaviours [72].

Conclusion

This study provides insights into how interviewers make professional judgments during high-stakes MMI selection processes. The findings underscore that judgment in MMIs is not purely objective but influenced by contextual factors, dialogue, and reflective practice. This research contributes to the broader assessment literature by linking the theory of work-based assessment to MMI practice, offering implications for improving the design, training, and implementation of MMIs in postgraduate selection. Importantly, the study also highlights the professional growth experienced by interviewers through their participation, suggesting that MMI selection can function not only as an assessment tool but also as a developmental process for interviewers themselves. These insights have the potential to enhance fairness, validity, and the overall effectiveness of future MMI assessments.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-024-06535-3.

Supplementary Material 1

Acknowledgements

We wish to thank Marcia Grant for her contributions to an earlier version of the article.

Author contributions

CR conceived of the study, and contributed to data collection, data analysis, and wrote the first draft. AB and KM contributed to the study design, data collection, data analysis and interpretation. KK contributed to data analysis and interpretation. All contributed to iterative revisioning and read and approved the final manuscript for submission.

Funding

The data contained in this publication was sourced from the Commonwealth funded evaluation activities for the Australian General Practice training program undertaken by the University of Sydney. The study design presented in this paper was conducted by the authors of this paper and approved by The University of Sydney Human Research Ethics Committee. The data collection, data analysis, interpretation of the data and writing of the manuscript was also completed by the authors, and not in collaboration with GPET.

Data availability

Datasets supporting the conclusions of this article are included within the article. The datasets generated and analysed during the current study are not publicly available due to confidentiality agreements approved by the Human Research Ethics Committee but are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All research method were conducted in accordance with relevant guidelines and regulations. The University of Sydney Human Research Ethics Committee approved the research. Protocol Number 13859. Informed written consent for participation was obtained from participants who were interviewed to enable us to include their data from this study. All participants were reassured that data were strictly de-identified to protect participant privacy.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Medicine and Population Health, Division of Medicine, The University of Sheffield, Sheffield, UK
[2]Sydney Medical School – Education Office, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia
[3]Sydney Medical School – Northern Clinical School, The University of Sydney, Sydney, NSW, Australia
[4]Division of Learning and Teaching, Charles Sturt University, Bathurst, NSW, Australia
[5]College of Medicine and Public Health, Flinders University, Adelaide, SA, Australia

Received: 30 January 2024 / Accepted: 16 December 2024
Published online: 03 January 2025

References

1. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Education: Theory Pract. 2013;18(2):291–303.
2. Wood TJ. Exploring the role of first impressions in rater-based assessments. Adv Health Sci Education: Theory Pract. 2013;19(3):409–27.
3. Govaerts MJ, Van de Wiel MW, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: raters' performance theories and constructs. Adv Health Sci Education: Theory Pract. 2013;18(3):375–96.
4. Berendonk C, Stalmeijer R, Schuwirth LT. Expertise in performance assessment: assessors' perspectives. Adv Health Sci Educ. 2013;18(4):559–71.
5. 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.
6. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box'differently: assessor cognition from three research perspectives. Med Educ. 2014;48(11):1055–68.
7. Dowell J, Lynch B, Till H, Kumwenda B, Husbands A. The multiple mini-interview in the U.K. context: 3 years of experience at Dundee. Med Teach. 2012;34(4):297–304.
8. Roberts C, Walton M, Rothnie I, Crossley J, Lyon P, Kumar K, Tiller D. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. Med Educ. 2008;42(4):396–404.
9. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. Med Educ. 2004;38(3):314–26.
10. Roberts C, Clark T, Burgess A, Frommer M, Grant M, Mossman K. The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. BMC Med Educ. 2014;14(1):169.
11. Patterson F, Rowett E, Hale R, Grant M, Roberts C, Cousans F, Martin S. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. BMC Med Educ. 2016;16(1):87.
12. Roberts C, Khanna P, Rigby L, Bartle E, Llewellyn A, Gustavs J, Newton L, Newcombe JP, Davies M, Thistlethwaite J. Utility of selection methods for specialist medical training: a BEME (best evidence medical education) systematic review: BEME guide 45. Med Teach. 2018;40(1):3–19.
13. Lundh A, Skjelsager K, Wildgaard K. Use of professional profiles in applications for specialist training positions. Dan Med J. 2013;60(7):A4648–4648.
14. Dore KL, Kreuger S, Ladhani M, Rolfson D, Kurtz D, Kulasegaram K, Cullimore AJ, Norman GR, Eva KW, Bates S et al. The Reliability and Acceptability of the Multiple Mini-Interview as a Selection Instrument for Postgraduate Admissions. *Academic Medicine* 2010, 85(10) Supplement(RIME):Proceedings of the Forty-Ninth Annual Conference November 7-November 10, 2010:S2060-S2063.
15. Hofmeister M, Lockyer J, Crutcher R. The multiple mini-interview for selection of international medical graduates into family medicine residency education. Med Educ. 2009;43(6):573–9.
16. Burgess A, Roberts C, Clark T, Mossman K. The social validity of a national assessment centre for selection into general practice training. BMC Med Educ. 2014;14(1):261.
17. Yoshimura H, Kitazono H, Fujitani S, Machi J, Saiki T, Suzuki Y, Ponnamperuma G. Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: a reliability and acceptability comparison. BMC Med Educ. 2015;15:75.
18. Sklar MC, Eskander A, Dore K, Witterick IJ. Comparing the traditional and multiple Mini interviews in the selection of post-graduate medical trainees. Can Med Educ J. 2015;6(2):e6.
19. Yamada T, Sato J, Yoshimura H, Okubo T, Hiraoka E, Shiga T, Kubota T, Fujitani S, Machi J, Ban N. Reliability and acceptability of six station multiple mini-interviews: past-behavioural versus situational questions in postgraduate medical admission. BMC Med Educ. 2017;17:1–7.
20. Axelson RD, Kreiter CD. Rater and occasion impacts on the reliability of pre-admission assessments. Med Educ. 2009;43(12):1198–202.
21. Baker KD, Sabo RT, Rawls M, Feldman M, Santen SA. Versatility in multiple mini-interview implementation: rater background does not significantly influence assessment scoring. Med Teach. 2020;42(4):411–5.
22. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? Med Educ. 2010;44(7):690–8.
23. Christensen MK, Lykkegaard E, Lund O, O'Neill LD. Qualitative analysis of MMI raters' scorings of medical school candidates: A matter of taste? Adv Health Sci Educ. 2018;23(2):289–310.
24. Alweis RL, Fitzpatrick C, Donato AA. Rater Perceptions of Bias Using the Multiple Mini-interview Format: a Qualitative Study, vol. 3; 2015.
25. Breil SM, Forthmann B, Hertel-Waszak A, Ahrens H, Brouwer B, Schönefeld E, Marschall B, Back MD. Construct validity of multiple mini interviews–investigating the role of stations, skills, and raters using bayesian G-theory. Med Teach. 2020;42(2):164–71.
26. 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.
27. Fung BSC, Gawad N, Rosenzveig A, Raîche I. The effect of assessor professional background on interview evaluation during residency selection: a mixed-methods study. Am J Surg. 2022;225(2):260–5.
28. Klusmann D, Knorr M, Hampe W. Exploring the relationships between first impressions and MMI ratings: a pilot study. Adv Health Sci Educ. 2023;28(2):519–36.
29. Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, Norcini J. I'm pickin' up good regressions': the governance of generalisability analyses. Med Educ. 2007;41(10):926–34.
30. Bégin P, Gagnon R, Leduc J-M, Paradis B, Renaud J-S, Beauchamp J, Rioux R, Carrier M-P, Hudon C, Vautour M. Accuracy of rating scale interval values used in multiple mini-interviews: a mixed methods study. Adv Health Sci Educ. 2021;26(1):37–51.
31. Tavares W, Kuper A, Kulasegaram K, Whitehead C. The compatibility principle: on philosophies in the assessment of clinical competence. Adv Health Sci Educ. 2020;25(4):1003–18.
32. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. Med Educ. 2011;45(10):1048–60.
33. Govaerts MJ, Van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. Adv Health Sci Educ. 2007;12(2):239–60.
34. Govaerts M, Schuwirth L, Van der Vleuten C, Muijtjens A. Workplace-based assessment: effects of rater expertise. Adv Health Sci Educ. 2011;16(2):151–65.
35. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011;86(10):S1–7.
36. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45(6):560–9.
37. Sebok SS, Syer MD. Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. Acad Med. 2015;90(11):S50–5.
38. Valentine N, Durning S, Shanahan EM, Schuwirth L. Fairness in human judgement in assessment: a hermeneutic literature review and conceptual framework. Adv Health Sci Education: Theory Pract. 2021;26(2):713–38.
39. Vleuten CP. When I say… context specificity. Med Educ. 2014;48(3):234–5.
40. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 2013;35(7):564–8.
41. Tavares W, Kinnear B, Schumacher DJ, Forte M. Rater training re-imagined for work-based assessment in medical education. Adv Health Sci Educ. 2023;28(5):1697–709.

42. Wong WYA, Roberts C, Thistlethwaite J. Impact of structured feedback on examiner judgements in objective structured clinical examinations (OSCEs) using generalisability theory. Health Professions Educ. 2020;6(2):271–81.
43. Crossley JGM, Groves J, Croke D, Brennan PA. Examiner training: a study of examiners making sense of norm-referenced feedback. Med Teach. 2019;41(7):787–94.
44. Walby K. On the social relations of research: a critical assessment of institutional ethnography. Qualitative Inq. 2007;13(7):1008–30.
45. DeVault ML. Introduction: what is institutional ethnography. Soc Probs. 2006;53:294.
46. McCoy L, Devault M. Institutional Ethnography: Using Interview to Investigate Ruling Relations. I: DE Smith (red.): Institutional Ethnography as Practice. In.: Oxford: Rowman & LittleField Publishers; 2006.
47. Burgess A, Roberts C, Sureshkumar P, Mossman K. Multiple mini interview (MMI) for general practice training selection in Australia: interviewers' motivation. BMC Med Educ. 2018;18(1):1–8.
48. Reeves S, Peller J, Goldman J, Kitto S. Ethnography in qualitative educational research: AMEE Guide 80. Med Teach. 2013;35(8):e1365–79.
49. Pope C. Conducting ethnography in medical settings. Med Educ. 2005;39(12):1180–7.
50. Pope C, Smith A, Goodwin D, Mort M. Passing on tacit knowledge in anaesthesia: a qualitative study. Med Educ. 2003;37(7):650–5.
51. Woolgar S. Knowledge and reflexivity: new frontiers in the sociology of knowledge. London [etc.]: Sage; 1988.
52. Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: problematising thematic emergence, triangulation, saturation and member checking. Med Educ. 2017;51(1):40–50.
53. Barry CA, Britten N, Barber N, Bradley C, Stevenson F. Using reflexivity to optimize teamwork in qualitative research. Qual Health Res. 1999;9(1):26–44.
54. McWhinney IR. Being a general practitioner: what it means. Eur J Gen Pract. 2000;6(4):135–9.
55. Roberts C, Wilkinson TJ, Norcini J, Patterson F, Hodges BD. The intersection of assessment, selection and professionalism in the service of patient care. Med Teach. 2019;41(3):243–8.
56. Patterson F, Roberts C, Hanson MD, Hampe W, Eva K, Ponnamperuma G, Magzoub M, Tekian A, Cleland J. 2018 Ottawa consensus statement: selection and recruitment to the healthcare professions. Med Teach. 2018;40(11):1091–101.
57. Crossley J, Jolly B. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. Med Educ. 2012;46(1):28–37.
58. Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently. Adv Health Sci Educ. 2013;18(3):325–41.
59. Wood TJ, Daniels VJ, Pugh D, Touchie C, Halman S, Humphrey-Murto S. Implicit versus explicit first impressions in performance-based assessment: will raters overcome their first impressions when learner performance changes? Adv Health Sci Educ 2023:1–14.
60. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. Med Educ. 2013;47(12):1164–74.
61. Coles C. Developing professional judgment. J Continuing Educ Health Professions. 2002;22(1):3–10.
62. Hammersley M. Ethnography: problems and prospects. Ethnography Educ. 2006;1(1):3–14.
63. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies:guided by Information Power. Qual Health Res. 2016;26(13):1753–60.
64. Roberts C, Kumar K, Finn G. Navigating the qualitative manuscript writing process: some tips for authors and reviewers. BMC Med Educ. 2020;20(1):439.
65. Hodges B, Paul R, Ginsburg S, the Ottawa Consensus Group M. Assessment of professionalism: from where have we come – to where are we going? An update from the Ottawa Consensus Group on the assessment of professionalism. Med Teach. 2019;41(3):249–55.
66. Bok HGJ, Teunissen PW, Favier RP, Rietbroek NJ, Theyse LFH, Brommer H, Haarhuis JCM, van Beukelen P, van der Vleuten CPM, Jaarsma DADC. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ. 2013;13(1):123.
67. Yeates P, Cope N, Hawarden A, Bradshaw H, McCray G, Homer M. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. Med Educ. 2019;53(3):250–63.
68. Munro D, Bore M, Powis D. Personality determinants of success in medical school and beyond:steady, sane and nice. Personality down under: Perspect Australia 2008:103–12.
69. Heeneman S, de Jong LH, Dawson LJ, Wilkinson TJ, Ryan A, Tait GR, Rice N, Torre D, Freeman A, van der Vleuten CP. Ottawa 2020 consensus statement for programmatic assessment–1. Agreement on the principles. Med Teach. 2021;43(10):1–10.
70. Roberts C, Khanna P, Lane A, Reimann P, Schuwirth L. Exploring complexities in the reform of assessment practice: a critical realist perspective. Adv Health Sci Educ. 2021;26(5):1641–57.
71. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85.
72. Rees CE, Ottrey E, Barton P, Dix S, Griffiths D, Sarkar M, Brooks I. Materials matter: understanding the importance of sociomaterial assemblages for OSCE candidate performance. Med Educ. 2021;55(8):961–71.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.