RUNNING HEAD: MMI raters’ scorings: A matter of taste?

Title:

Qualitative analysis of MMI raters’ scorings of medical school candidates: A matter of taste?

Authors:

Mette K. Christensen*, Eva Lykkegaard*, Ole Lund*, Lotte D. O’Neill**

Affiliation(s):

* Centre for Health Sciences Education, Aarhus University

** SDU Centre for Teaching and Learning, University of Southern Denmark

Full address for correspondence:

Mette Krogh Christensen

Centre for Health Sciences Education, Aarhus University, Palle Juul-Jensens Boulevard 82, 8200 Aarhus N, Denmark

Telephone: +45 86205227 / Mobile: +45 31343093 / E-mail address: mkc@cesu.au.dk
ABSTRACT

Recent years have seen leading medical educationalists repeatedly call for a paradigm shift in the way we view, value and use subjectivity in assessment. The argument is that subjective expert raters generally bring desired quality, not just noise, to performance evaluations. While several reviews document the psychometric qualities of the Multiple Mini-Interview (MMI), we currently lack qualitative studies examining what we can learn from MMI raters’ subjectivity. The present qualitative study therefore investigates rater subjectivity or taste in MMI selection interview. Taste (Bourdieu, 1984) is a practical sense, which makes it possible at a pre-reflective level to apply ‘invisible’ or ‘tacit’ categories of perception for distinguishing between good and bad. The study draws on data from explorative in-depth interviews with 12 purposefully selected MMI raters. We find that MMI raters spontaneously applied subjective criteria – their taste – enabling them to assess the candidates’ interpersonal attributes and to predict the candidates’ potential. In addition, MMI raters seemed to share a taste for certain qualities in the candidates (e.g. reflectivity, resilience, empathy, contact, likeness, ‘the good colleague’); hence, taste may be the result of an ongoing enculturation in medical education and healthcare systems. This study suggests that taste is an inevitable condition in the assessment of students’ performance. The MMI set-up should therefore make room for MMI raters’ taste and their connoisseurship, i.e. their ability to taste, to improve the quality of their assessment of medical school candidates.

KEY WORDS

Admission; Assessment; Bourdieu; Medical school candidates; MMI rater’s scorings; Multiple Mini-Interview; Qualitative study; Selection; Subjectivity; Taste;
Running head:

MMI raters’ scorings: A matter of taste?

Title:

Qualitative analysis of MMI raters’ scorings of medical school candidates: A matter of taste?

Authors:

Mette K. Christensen*, Eva Lykkegaard*, Ole Lund*, Lotte D. O’Neill**

Affiliation(s):

* Centre for Health Sciences Education, Aarhus University

** SDU Centre for Teaching and Learning, University of Southern Denmark

Full address for correspondence:

Mette Krogh Christensen

Centre for Health Sciences Education, Aarhus University, Palle Juul-Jensens Boulevard 82, 8200 Aarhus N, Denmark

Telephone: +45 86205227 / Mobile: +45 31343093 / E-mail address: mkc@cesu.au.dk
ABSTRACT

Recent years have seen leading medical educationalists repeatedly call for a paradigm shift in the way we view, value and use subjectivity in assessment. The argument is that subjective expert raters generally bring desired quality, not just noise, to performance evaluations. While several reviews document the psychometric qualities of the Multiple Mini-Interview (MMI), we currently lack qualitative studies examining what we can learn from MMI raters’ subjectivity. The present qualitative study therefore investigates rater subjectivity or taste in MMI selection interview. Taste (Bourdieu, 1984) is a practical sense, which makes it possible at a pre-reflective level to apply ‘invisible’ or ‘tacit’ categories of perception for distinguishing between good and bad. The study draws on data from explorative in-depth interviews with 12 purposefully selected MMI raters. We find that MMI raters spontaneously applied subjective criteria – their taste – enabling them to assess the candidates’ interpersonal attributes and to predict the candidates’ potential. In addition, MMI raters seemed to share a taste for certain qualities in the candidates (e.g. reflectivity, resilience, empathy, contact, alikeness, ‘the good colleague’); hence, taste may be the result of an ongoing enculturation in medical education and healthcare systems. This study suggests that taste is an inevitable condition in the assessment of students’ performance. The MMI set-up should therefore make room for MMI raters’ taste and their connoisseurship, i.e. their ability to taste, to improve the quality of their assessment of medical school candidates.

KEY WORDS

Admission; Assessment; Bourdieu; Medical school candidates; MMI rater’s scorings; Multiple Mini-Interview; Qualitative study; Selection; Subjectivity; Taste
INTRODUCTION

In recent years, leading medical educationalists have repeatedly called for a paradigm shift in the way we view, value and use subjectivity in assessment. It has been questioned whether what has traditionally been considered ‘noise’ and undesirable rater subjectivity from a measurement perspective is not precisely what simultaneously contains and adds desirable professional expertise that is fundamental to quality judgments (van der Vleuten et al., 2012; Govaerts and van der Vleuten, 2013; Hodges, B., 2013; van der Vleuten, 2014). Van der Vleuten (2014, p. 235) described his standpoint as follows:

“*The new understanding that we are currently documenting is that assessors are not passive, perfectly calibrated measurement instruments, but active agents constructing judgements using information from situational factors and personal experiences. From this perspective, different assessors are not expected to make similar judgements, and this may actually be desirable! Indeed, from a learning orientation, different perspectives on performance may be beneficial. Also, we are moving away from scoring by numbers. Complex skills and performances are elusive and need the wealth of information that only narrative can provide. Perhaps we should invoke qualitative methodologies to give meaning to our interpretations.*”

Moreover, the need to assess complex skills and performances has been increasing. Recent decades have seen the advent of a number of national frameworks describing desirable ‘non-cognitive competences’ such as inter-personal skills for medical students and doctors. Simultaneously, accumulating evidence in the field of student selection indicates that the traditional panel interview is not a particularly valid or reliable test format for the assessment of interpersonal qualities upon admission to medical education (Salvatori 2001). This has triggered the development of the Multiple Mini-Interview (MMI) format, which has become a preferred admission interview format at many medical schools and in a number of countries (Andreassen et al., 2016; Kumar et al., 2009; Pau et al., 2013). There appears to be
considerable local variety in the non-cognitive MMI test domains. The following are just some of
the desirable traits reported; communication, giving and receiving instructions, ethical or moral
reasoning, resilience, maturity, enthusiasm, initiative, integrity, responsibility, handling stress,
tolerance for ambiguity, collegiality, teamwork advocacy, cultural sensitivity, empathy, the ability
to self-assess, and self-awareness, etc. (Eva et al., 2004b; Harris and Owen, 2007, Lemay et al,
2007; Gafni et al., 2012; Dowell et al., 2012). The MMI format was developed in order to overcome
the general phenomenon of ‘context specificity’ of behaviour by ensuring sufficient sampling of
situational contents and raters (Eva et al., 2004b; van der Veluten, 2014). Studies rooted in
Generalizability Theory have since repeatedly confirmed that station content is more often a
larger source of score variance than rater biases are (Eva et al, 2004b; Axelson & Kreiter, 2009;
Onyon, Wall and Goodyear, 2009; Hofmeister, Lockyer and Crutcher, 2009; Seebok, Luu and
Klinger, 2015; Knorr and Hissbach, 2014). It is therefore usually recommended to sample more
stations containing different test contents rather than using more raters per station (Knorr and
Hissbach, 2014). Generalizability Theory (and Classical Test Theory) breaks down the observed
score variance into ‘true score’ variance, i.e. the variance in scores attributable to true/real
differences in test takers’ competencies, and ‘error’ variances of different sorts, including raters’
interactions with the contents and the test takers (Brennan, 2001). The underlying validity
assumption is typically that a specific trait (e.g. applicants’ communication skills) is being tested
in a group of test takers. From that perspective, it is obviously highly undesirable if test scores
mainly reflect some rater characteristic. In the seminal MMI study by Eva and colleagues (2004a),
the score variance reflecting raters’ candidate-specific subjectivity (i.e. taste) was estimated to
reach 45%, a circumstance, which could be ameliorated by adequate sampling. The innovators
who developed the MMI (Eva et al., 2004b) concluded that rater subjectivity did not preclude
reliable and valid assessment of individual’s abilities.
Subsequently, a number of reviews have documented the psychometric qualities of the MMI (Pau et al., 2013; Knorr and Hissbach, 2014; Patterson et al., 2016; Rees et al., 2016), but there is currently a lack of studies examining what we can learn from MMI raters’ subjectivity from a qualitative perspective (Patterson et al., 2016). We found only one qualitative study examining MMI raters’ experiences, but we agree with Kumar and colleagues (2009) that qualitative studies of performance assessments have added value relative to quantitative studies, and that they may enrich our understanding of the challenges faced.

Given the existing tensions in the field regarding subjectivity as adding both noise and desirable professional expertise to assessment simultaneously, and the fact that we found no qualitative studies of how MMI rater subjectivity operates, it seemed important to explore MMI rater subjectivity further.

Theoretical framework

The theoretical framework in this study applied the term ‘taste’ in line with the sociological theory of Pierre Bourdieu (Bourdieu, 1984; Bourdieu, 1998) and Nick Crossley (Crossley, 2013).

Normally, taste relates to the sensory impressions we get from manipulating and dissolving food in our mouth and sensing its flavours with our taste buds. However, taste can also be related to subjectivity and the individual’s practical sense (Bourdieu, 1998, p.25) of a situation:

“...practical sense, that is, an acquired system of preferences, of principles of vision and division (what is usually called taste), and also a system of durable cognitive structures (which are essentially the product of the internalization of objective structures) and of schemes of action which orient the perception of the situation and the appropriate response”.

The concept of taste emphasises how humans’ way of making sense of or assessing the world is primarily a matter of perception rather than intellectualisation. Bourdieu contends that taste is a practical mastery of distribution (Bourdieu, 1984), which makes it possible at a prereflective level
to apply ‘invisible’ or ‘tacit’ categories of perception when we distinguish between good and bad. One might therefore assume that taste is an individual capacity. However, Bourdieu argued that tasting, i.e. judging and assessing the world, is the product of incorporated and habituated fundamental structures of a society, and that the incorporation of these shared societal structures leads to classificatory schemes (Bourdieu, 1984) by which we classify and categorise cultural forms (Nick Crossley, 2013) such as art, food, academic or sports performances (Christensen, 2009). In short, taste is formed by social interactions and vice versa. Sociologist Nick Crossley (2013) described the interpersonal characteristics in the forming of taste: so-called alters (for example role-models, significant others or authorities) are a source of exposure to cultural forms that actors (in this case the MMI raters) might not otherwise encounter. Alters teach actors how to appreciate and enjoy cultural objects that they might not otherwise “get”. Actors are motivated to make an effort to learn to like what alters like, because actors desire to be recognised by alters and to share/consume the cultural forms with others. Actors may consequently develop positive associations (or taste) to cultural forms. However, distaste for the cultural forms may create “cognitive dissonance” for the actor, which he/she resolves by trying to like those forms — and in some cases succeeding in doing so (Crossley, 2013). Accordingly, in relation to this study, we may define taste as an incorporated reproduction of a shared habitus by which certain cultural forms in medical education are assessed as more qualifying than others or, in other words, we may say that certain performances taste better than others do.

When it comes to classifying and categorising medical school candidates as qualified or not, the MMI raters’ classificatory schemes may be assumed comparable because the raters are actors of the same society in general and of the same social field, i.e. medical education, in particular. Even so, their classificatory schemes may also differ because the raters are actors in different sub-fields within medical education. Concerning the MMI raters’ assessment as a matter of taste, we may assume that the raters are prone to seek out candidates with whom they share tastes.
We apply the concept of taste as the theoretical framework for our investigation because it pinpoints how judgements – in this case assessments of candidates' performance at the MMI – are based not only on logic and rationality, but also context-specificity, practical experiences, and intuitions. Taking this starting point will enable us to renew the discussion on how raters' subjectivity operates during the MMI test.

The aim of this paper was to qualitatively investigate MMI raters' subjectivity with regard to their taste (Bourdieu, 1984) for medical education candidates. Taste is a sociological concept describing individuals' subjective judgments as a matter of practical sense. The research questions were:

- How do MMI raters experience taste operating during the MMI and, in particular, how do they experience the interaction between objective criteria and taste?
- What characterises MMI raters' taste?
- Which classificatory schemes underpin MMI raters' taste?

**METHODS**

Tasting functions below the level of immediate discourse (Bourdieu, 1984). Consequently, to gain insight into the 'invisible' classificatory schemes that guide MMI raters' taste and assessment, we use a constructivist orientated approach where the presented meanings, experiences, and orientations are con-structed among the informant, the interviewer, and the research team (Charmaz, 2006). This approach explicitly treats researchers' works as constructions and not objectified products (Charmaz, 2000). We applied a qualitative, explorative in-depth interview method with focus on twelve MMI raters' narrated experiences of assessing medical school candidates during MMI tests. We base the choice of this approach on its sensitivity to the informants' experiences (Garro and Mattingly, 2000) and to their descriptions of individual preferences (Christensen, 2009).
The context of the study

The study context was the test-based admission track to Medical School, Aarhus University (Denmark). The first step in the admission track was a requirement of minimally acceptable upper secondary grades. The second step was participation in a multiple choice exam testing generic reasoning and thinking across the two broad domains of mathematics/science and humanities/social sciences (the uniTEST developed by the Australian Council of Educational Research). The third and final step was participation in an MMI developed to test ‘non-cognitive’ competences. The MMI was developed and pilot tested by the Centre for Health Sciences Education, Aarhus University (Denmark) in cooperation with invited stakeholders representing regional clinicians, lecturers, patient organisations and the national Danish medical association. Four test domains (communication, empathy, collaboration and resilience) were identified with the nominal group technique, and 8 different MMI stations (2 per test domain) were developed and pilot tested (Andreassen et al., 2016). Ahead of the test day, raters participated in a 2-hour training session, which contained general practical information and background information about scoring, scales, and some scoring practice. Furthermore, raters were invited to participate in this study. Subsequently, 175 applicants and 104 MMI raters participated in the MMI in June 2016. Some MMI stations were occupied by one rater, others by two raters. Each rater rated candidates on a 7-point Likert scale, which contained verbal descriptions of examples of observed behaviours corresponding to the scores 1 (unsatisfactory), 4 (acceptable), and 7 (excellent) for each of the two non-cognitive domains tested on the station. In addition, they also rated applicants on an overall ‘suitability for Medicine’ scale, which was a similar but unanchored scale. No practically applicable scoring rubrics or anchors are ever exhaustive or complete in authentic test situations of human performance. During the training session before the MMI, raters were therefore specifically advised to also embrace and use their subjective expertise, to employ the whole range of the suitability scale and to go with their best estimate of a judgment, just as they would when participating as raters in any job interview.
Participants

The participants were a convenience sample of volunteer raters. During the abovementioned 2-hour training session, raters were briefed about this research project and invited to participate. Of the 104 MMI raters, 40 volunteered to participate in qualitative interviews and therefore shared their e-mail addresses and/or phone numbers with us. The participant selection favoured raters who were able to participate in the in-depth interview within a week from the MMI test - we assumed this would strengthen the trustworthiness of the study. The participants were additionally sought to best represent the group of rater experiences and therefore all four test domains were represented among the included participants, as we assumed this would strengthen the richness of the study. Moreover, MMI raters who gave low scores to at least one MMI applicant (1 or 2 on the 7-point Likert scale) were purposively selected for this study, as we assumed that these MMI raters would better remember their taste during the rating of low-scoring candidates.

Before the MMI session, we identified 24 possible participants of the 40 volunteer raters based on the above mentioned criteria. Immediately after the MMI session, 12 of those participants were selected for interview appointments (see Table 1 and 2). Guided by the concept of information power (Malterud et. al, 2015) we assess that the highly specific participant group for this study aim (12 of 104 raters who participated in the MMI not more than a week ago), the use of Bourdieu's (established) theoretical concept of taste and a high quality interview dialogue (performed by two experienced interviewers) makes 12 raters a comprehensive number of participants to answer our research question.

The interviews were arranged using the contact information and scheduled for the same week. Before participating in the study, all participants were informed about the study verbally and in writing, and they signed a written statement of consent. The participants were anonymised in the presentation of data.
The authors obtained permission from the Danish Data Protection Agency to use and combine the specific data generated in interviews for the purpose of this study, as required by Danish law. In concordance with the guidelines of the Regional Ethics Committee in Denmark, the study was exempt from ethical approval because of its qualitative design.

**Data collection**

A semi-structured interview guide (Kvale and Brinkmann, 2009) was prepared and then applied in each of the interviews. In the interview, the participants were invited to exemplify what during the MMI characterises the performances of ‘good’ and ‘bad’ medical school candidates, respectively. They were asked to describe one or more candidate(s) they had rated low during the MMI, and encouraged to reflect upon the justice in using their subjective experiences during their ratings. In addition, they were asked to describe and exemplify what characterizes a good medical student and a good physician, and they were asked to emphasise any attributes or role models that may have inspired them when they were students or young physician themselves.

The purpose of these questions was to investigate which preferences in taste they may have used during the MMI interviews.

The volunteering MMI raters were interviewed individually by the first or the second author in the week following the MMI. Both authors were experienced qualitative interviewers. The interviews were audio recorded and lasted 38-59 minutes (mean: 49 minutes).

**Data analysis**

In order to organise and categorise the data, we developed a matrix (Miles, Huberman and Saldaña, 2014). The matrix was created as a folder in Excel. The questions of the interview guide served as predetermined coding categories and thus appeared as 27 column labels in the matrix. The labels of the twelve rows contained the MMI raters’ names. The second and third author listened to the interviews and coded each interview segment-by-segment (Charmaz, 2006) by adding quotations (segments) from the MMI raters' interviews to the cells of the 27 coding
categories, thus applying a meaning categorisation strategy (Kvale and Brinkmann, 2009). The quotations that were added to the matrix were noted in a format that came as close to what was actually said as possible. Moreover, every quotation was assigned a time code specifying when it occurred during the interview. The purpose of this procedure was to remain descriptive in the initial phase of the analysis, to provide a preliminary overview of what each participant stated in relation to the interview questions and to facilitate the following cross-case analyses. In this initial phase, new coding categories would surface and be added as new columns to the matrix. In total, ten new coding categories were added.

Next, the third author condensed (Kvale and Brinkmann, 2009) the categorised quotations by creating essence-capturing phrases. The purpose was to get to the core of what was stated, but also to create connections in between the different MMI raters’ experiences as well as capturing the breadth and diversity of the data. Categories and essence-capturing phrases were discussed and interpreted at regular meetings between all authors until agreement was reached. In this phase, we gradually drew on the sociological theory of Pierre Bourdieu and Nick Crossley in order to reach a collective understanding of the data.

As a consequence of the insights thus gained during the first and second steps of the analysis, we proceeded to the third and final step where we conducted a theoretical reading (Kvale and Brinkmann, 2009) based on the two sociological concepts of taste and classificatory schemes, as already described. We applied these concepts to facilitate theoretically informed interpretations of 1) how taste operates during the MMI, 2) the characteristics of MMI raters’ taste and 3) the classificatory schemes underpinning the MMI raters’ taste.

RESULTS

The results are presented as answers to the three research questions.
How do MMI raters experience taste operating during the MMI test?

The first research question concerned the way the raters experienced their own use of taste during the MMI interviews, including the dilemmas that they faced in the process.

The raters were asked if it was easy for them to assess the candidates' suitability for medical school:

Female rater 7: I actually thought it was relatively easy. It came naturally to me.

Male rater 3: You weren't in doubt at all. They had spoken for less than two minutes and you just knew that this one needs to be enrolled for Medicine [...], whereas all of the mid-range applicants are harder to assess.

The data indicates that the taste of the raters allowed them to assess the candidates' performance effortlessly. Assessing the candidates came naturally and quickly to them. This was further underlined by a rater who told how she became so caught up in the narratives of the candidates that she forgot to use the objective criteria (the anchored scales): "At times I get caught up [in the applicant's story] and forget to send out the "objective doctor" to inspect!" (Female rater 5). This might also imply that while the MMI raters spontaneously apply subjective criteria, i.e. their taste, in the assessments of candidates, they may have to struggle to stay analytical and judge the candidates in accordance with standardised criteria and the anchored scales. This seems to underline the theoretical assumption that taste is fundamental to any kind of human judgement and therefore cannot be bypassed by any attempt to objectify the testing process.

Most raters underlined that good candidate performances could not simply be reduced to the contents they were conveying and their ability to articulate what the anchors described. As exemplified below, some candidates were saying specific things which, objectively judged, ought to result in high scores; however, the MMI raters' practical senses of the candidates overruled such an objective assessment:
Male rater 3: She seemed both intelligent and bright and said the right things. But she came across as very rigid and failed to look me into the eyes throughout [...] 

Female rater 2: Another student provided many good points, but her face was completely neutral. [...] And we took the liberty of asking the actor afterwards [a roleplay station] how she had experienced the situation; and she felt as if there had been no contact with the person. So I also scored relatively low.

The candidates were also judged on their abilities to connect and establish a sense of interpersonal contact with the rater or others who took part in the particular station. In connection with this, the raters recounted being attentive to the candidates' non-verbal and bodily way of interacting during the face-to-face encounter of the test. Especially, they seemed to focus on the candidates' eye movements and facial gestures, and the raters clearly expressed a taste for candidates who were able to maintain eye contact and perceived this as a sign of the candidates' openness and maturity.

The raters' tastes contributed to their assessment of the candidates with predictions that went beyond what was actually said and done during the test. As an example, one MMI rater (Male rater 4) commented on a candidate who had failed to meet the last criterion on the anchor. The rater nevertheless felt that the candidate would have managed to do so if he had been given more time, because he did not seem to be bothered by talking about private and sensitive issues with the patient (an actor at the MMI station). Therefore, the rater awarded the candidate the highest possible score. Another MMI rater explained:

Male rater 1: I remember one person who scored 2 or 3 [on the anchored scales], and I thought, that's immaturity. That person did, in fact, try, but only just scratched the surface and started talking about herself in a friend-to-friend kind of way [...] But even so, I thought she was trying and I had the feeling that [...] she will catch up. So I scored her relatively high [on the suitability scale]. In fact I might have used it [the suitability scale]
to counterbalance some of the ones that scored really low due to some measurable stuff, but where I thought that they would do OK anyway.

This MMI rater recounted that stringent use of the anchor criteria could paint a misleading picture of the candidate because it might not correspond with the raters’ overall sense of the candidate’s potential.

These statements indicate that – when relying on their taste – raters transcend what can be observed and measured objectively during the interview situation. The statements also show that they intuitively understand the candidates’ potential and how they may function as medical students or doctors in the future. Furthermore, the raters seemed to have developed this ability because they have first-hand experience with dealing with the practical demands of both medical school and medical practice, e.g. “I know because I have been in a similar situation” (Male rater 3). Also, one MMI rater exemplified that one of the things that guided her when discarding candidates during the MMI was her practical experiences with peers and colleagues who did not thrive in their job (Female rater 1).

In summary: Raters taste came naturally and quickly to them. In addition, they felt that their taste not only enabled them to assess the candidates’ interpersonal attributes in the situational context but also their future potential.

The relation between objective and subjective assessments

A sub-question of particular interest in the first research question was how raters experienced the interaction between objective criteria (the descriptive anchor examples provided in the MMI test) and taste. As mentioned earlier, the raters at each station were asked to assess the candidates according to two predefined attributes/domains (communication, empathy, cooperation or resilience) based on anchored Likert scales and to assess the candidates’ general suitability based solely on the raters’ subjective judgement (the overall scale). As already indicated, these assessments might have been mutually influencing:
Male rater 5: But it's all connected [all of the parameters]. I awarded some a higher general suitability score or conversely. But often scores were close.

The data indicates that the raters experience an intimate relationship between the predefined domains and subjective and objective assessments. This is further stressed by the quote mentioned earlier, which showed how a rater used subjective judgements to counterbalance his 'objective' assessments (Male rater 1). In contrast, another MMI rater (Female rater 1) explained that she would give some candidates lower scores than initially indicated by her gut-feeling, simply because these candidates failed to meet the anchor criteria. This suggests that the anchor criteria triggered doubts and that she therefore corrected her immediate tasting.

Correspondingly, the raters may have experienced that the instructions given to use the full range of the markings scale (Female rater 1) and to provide written explanations for every low score (Female rater 1; Female rater 3) affected the way they judged the candidates and occasionally made them reconsider their initial gut feelings.

A recurring theme in the interviews was the uncertainty expressed by the raters with respect to the accuracy of their own taste. Many raters spontaneously devalued intuitions as a valid assessment method, but still recognised the importance of their personal and practical experiences as a basis for their judgements:

Female rater 4: It is not hard for me to reach my own personal assessment, but whether it is expedient given what's out there, well that I can't really know. [...] I like the way I practice Medicine, but there are other doctors who act differently, but who are equally capable.

Male rater 2: It is difficult, because it's never been tested if I'm right [if my assessments are correct]. But I feel relatively confident that they are. [...] I have accepted that I can make mistakes.
Another rater (Female rater 5) experienced that it varied from rater to rater whether they placed more emphasis on their immediate intuitions based on taste or were more analytical in their approach. Furthermore, she questioned the stringency of the individual rater and whether the candidates were being measured on equal terms. This indicates that the raters did not experience their decision-making processes as being fully transparent, representative or reproducible in any positivistic sense because their assessments were inevitably affected by the particularities and context-specificity of the test situation and each rater’s intuition.

One of the reasons why the raters expressed uncertainty – and even discomfort – when assessing the candidates using their taste may be that they are encultured into a medical field that is pervaded by a scientific belief system that regards evidence-based knowledge – and not intuitions – as the highest of standards. This seemed to be confirmed by the fact that the raters felt comforted when told that the MMI is based on a multiple station design and that a large number of stations are supposed to neutralise rater subjectivity (Female rater 4; Female rater 1).

In summary: The relationship between subjective and objective assessment criteria was perceived by raters to be intimately intertwined, but even so there was an underlying uncertainty and discomfort associated with relying on subjectivity and taste.

What characterizes MMI raters’ taste?

The second research question concerned the characteristics of MMI raters’ taste.

The raters mentioned a wide range of characteristics, which we collected and unified under three overarching themes. Thus, we suggest that the raters show a taste for 1) candidates who displayed reflectivity and resilience, i.e. an ability to rise above themselves and the current task or situation and to embrace several aspects of the situation without being frustrated, 2) candidates who displayed contact and empathy, i.e. an ability to get in touch with others and the current task and situation, and 3) candidates who acted as ‘the good colleague’ and a certain
degree of alikeness with the rater, i.e. an attribute in the candidate that indicated ‘a good match’ or a ‘fusion of horizons’ with the rater.

**Reflectivity and resilience**

The raters assessed the candidates on their ability to rise above themselves, e.g. their personal intuitions and opinions and the situation in which they were involved:

Female rater 1: The girl who was awarded a 7 had a winning character and was sincerely empathic and very nuanced, and immediately – without any need for guiding questions – managed to involve all the persons [in the case task] really well. She also contributed with her personal perspective, but clearly stated that it was her perspective and that she also considered a broader social perspective.

The raters seemed to be on the look-out for virtues appreciated within academic or professionalised contexts such as being able to go beyond an intuitive and personal perspective and to see a situation from various perspectives. Some raters emphasised the candidates’ abilities to self-reflect, for instance being able to reflect on actions in the past and demonstrate that they had learned from such considerations (Female rater 4). Another rater highlighted two candidates who were able to establish an overview of and manage the tasks they were given at the station and to make ongoing adjustments to their performances together (Male rater 5). A third rater explained:

Female rater 2: And then there was the one who was capable of maintaining a structured conversation [which] we use a lot, particularly in my speciality when we need to confirm the consensus with the patient, because we often see many patients in a single day. So she achieved a really high score.

In contrast, the raters felt distaste for candidates who failed to rise above the conditions of the test situation, i.e. when being caught up in nervousness or frustration. The raters described how the MMI test was conducted in an atmosphere of anxiety. All candidates seemed to be influenced
by the gravity of the test, although some were better equipped to deal with this than others. Some candidates were almost paralysed and this influenced their assessment:

Male rater 2: The ones who achieved the poorest scores with me were nearly apathetic. Even though they were really only being asked to reflect on a situation, I noticed that they were – paralysed might be a bit harsh – but they were like: "uhm, I really don’t know what to answer". Which made me think [...] that they were not robust enough to get a grip, think things through and then give their best answer.

Thus, raters expressed a taste for the candidates who were able to keep their nerve (i.e. Male rater 5; Female rater 4). This included candidates who did not let themselves be stressed by the challenges set by the task or the 8-minute time frame (Female rater 1) and who took their time to read and consider the task before answering, and those who overcame a nervous start and got a hold of themselves (Male rater 4). This also suggests that performance was not assessed exclusively based on what the candidate did towards fulfilling anchor criteria such as ‘ability to include different perspectives of the case’ and ‘ability to self-reflect on own learning outcomes’, etc. The raters indicated that their assessment of resilience was also related to, for instance, the candidates’ balance or rootedness (Male rater 4), maturity (Male rater 3) or sincerity (Male rater 5), which seemed to be sensed from the interpersonal dynamics and sensory impressions that developed in the face-to-face encounter with the candidates.

**Contact and empathy**

While the raters seemed to value the candidates’ ability to keep situations at arm’s length, they also valued that the candidates demonstrated an ability to connect and empathise with others. One rater described this as a matter of being professionally distanced and compassionate at the same time (Male rater 4). Another rater (Female rate 2) emphasised this theme by describing a female candidate as follows:

Female rater 2: She really had the capacity to embrace the patient without saddening herself. She would say things like, "I do understand that" [...] and I was left with the feeling
that I wonder if she has started working as a nurse at a young age or something. She didn’t look very old, but definitely had learnt a thing or two and was used to talking with people.

In contrast, the raters experienced that low-scoring candidates were often unable to connect and empathise with others:

Male rater 1: I have met a great deal of students in the course of my career as a doctor, and I know how hard is to face those situations, so one needs to stay humble when it comes to patients. Therefore it triggers something in me to see fledglings that think they can do just about anything [...] And many of them asked closed questions and never got into the substance of the patient’s problem [...] And some didn’t listen but provided examples from their own life etc.: "Well, you simply need to...".

According to the raters, being in contact with patients and other participants on the stations was also a matter of displaying humility, responsiveness and tolerance, e.g. by showing an interest in the other and asking them open-ended questions. Furthermore, some raters also explained that they valued candidates who displayed passion for and invested themselves in the task at the station, for instance by telling a personal story that had affected them deeply. In contrast, candidates with low scores often struggled to identify with and invest themselves in the problem at the station. As an example, one rater (from the communication station) explained about a candidate who felt the need to turn to himself rather than to the patient (an actor at the MMI station) and ask: "Am I supposed to talk about this?" (Male rater 4).

The raters had an eye for candidates who displayed an ability to get in touch with others nonverbally. As earlier mentioned, eye contact seemed to be an important indicator for whether the candidates were capable of showing responsiveness and establishing inter-personal contact:

Female rater 3: If I talk with someone, I like them to look me in the eye. Naturally, your gaze may occasionally flicker [...], but most of the time I think they should look people in
the eye. I don't like when you have eye contact for a few seconds after which they start
looking elsewhere.

The importance of making eye contact and displaying responsiveness through body language was
even more evident in the descriptions that the raters gave of candidates who scored low. The
descriptions that the raters gave of candidates who scored low. The
element below illustrates the raters' taste and distaste for candidates' body language in
connection with inter-personal contact:

Female rater 2: And one of them [the candidates] simply threw himself back in the chair
like this [she shows how the candidate carelessly leaned back in the chair]. He just sat
there and at same time started [giggling] a bit. I just cannot take that seriously. If you sit
like this at the admission test for the education of your dreams when you are facing an
actor/patient who is close to bursting into tears... That's simply too relaxed for my taste.

The candidates who scored low displayed contact, manners and ways of interacting with others
that were immediately perceived as inappropriate or lowbrow because they were very different
from that of the raters and incongruent with the perceived needs of their daily practice. This
resulted in an immediate feeling of dissonance between the rater and these candidates, and the
raters' impression of the candidate instantly suffered.

Alieneness and 'the good colleague'

In contrast to the above, the raters seemed to be searching for alieneness, i.e. characteristics and
actions that allowed them to identify with the candidates:

Male rater 2: I think the ones I remember most clearly had the same analysis of the
situation as I had. And as we were supposed to be subjective, I simply assumed that my
assessment of the situation was the correct one.

The data exemplify the raters' taste for candidates who mirrored or even outperformed their
ways of performing during their career as medical students and physicians. In addition, the raters
valued candidates who displayed characteristics of 'the good colleague':

20
Female rater 3: It was much about if they would be a good colleague, and if I could envision them in the job [...] if I would want to take them with me as students and have them follow me around, and would I feel confident sending them into a patient alone knowing that I would sign and carry responsibility for whatever happened.

Some raters scored candidates who they perceived as being the opposite of a 'good colleague' very low (Female rater 6; Female rater 5). Others showed a distaste for candidates with an 'unprofessional' look, i.e. who had piercings and tattoos, and followed a particular clothing style, i.e. that they perceived as being excessively casual or low cut (Female rater 2).

The above data show that the raters had a distaste for unlikeness and a taste for alikeness or 'common sense', which in this context means sensing moments of shared sense-making, i.e. perceiving that the candidates' way of perceiving (and classifying) the given task and situation was in line with their own. Therefore, the raters’ ‘appetite’ for candidates seemed to grow the more the candidates were able to reveal or give them a foretaste of a practical sense similar to that of the raters or the raters' good colleagues.

In summary: Raters' tastes seemed to value applicant attributes in the realm of reflectivity, resilience, contact, empathy, being a good colleague and alikeness.

Which classificatory schemes underpin the MMI raters’ taste?

The third research question concerned the classificatory schemes underpinning MMI raters’ taste. The raters seemed to share a taste for certain attributes, which may be linked to a certain homogenisation and cultivation of shared habituated norms and values that occurred in the course of medical school (if not earlier). For example, one of the raters (Female rater 2) exemplified enculturation by the way the students dress and behave more and more alike during medical school, but it seemed to go deeper than this and to profoundly affect the students' way of being, thinking and making sense of the world. As exemplified by Male rater 3, who recounted he became less and less idealistic during medical school. Initially, he was ultra-liberal and interested
in social inequality; now he votes in the “middle of the road” and is interested in “geeky medical
stuff” (Male rater 3). This enculturation seemed to provide the raters with a shared backdrop – in
the form of classificatory schemata – for assessing the candidates, even though there may be
"nuances depending on the specialty” (Female rater 4) and the ward in question:

Female rater 1: I think it may be different from one place to the next as you are very much
shaped by the culture you are a part of [...] But yes, I think you are fairly agreed about
what is insufficient, medically speaking, and also what is not good enough with respect to
attitudinal challenges.

The profound effect that enculturation has had on the raters may have contributed to their taste
for alikeness and made candidates who displayed comparable cultural capitals (Bourdieu, 1998)
in their test performances immediately familiar and thus more attractive in the eyes of the raters.

The raters also told of challenging times when they were undergraduates. They faced an extensive
curriculum and challenging examinations. This meant not only that they had to develop their self-
and work-discipline to ensure that they studied regularly, but also that they took their studies
seriously and regarded studying as full-time work (Male rater 1). These experiences may have
been a significant source of the raters’ taste for reflectivity and resilience. Furthermore, the
workload meant that the raters had experienced the need to establish a study-leisure time
balance (or work-life balance) to avoid burn-out, e.g. it was necessary not to let oneself become
excessively stressed by the workload and learn that it did not help one’s understanding/learning
to study excessively (Female rater 4). In this way, the raters also seemed to develop a pragmatic
relationship to their studies, which appeared to be further habituated by meeting medical
practice. The raters explained how practice taught them that the world is more complex than it
appears to be in the books:

Male rater 4: At medical school they test if you know what’s on page 117, and as a doctor
it is more about the real world [...] there, patients have several conditions and they do not
always meet all the criteria.
The raters described how the complexities of medical practice had forced them to take a more pragmatic approach towards textbook knowledge, which indicates that they underwent a development from understanding their textbooks as manuals for how to deal with reality to understanding and being concerned with the importance of contextual practicalities, such as dealing with practical dilemmas, conflicting interests or identifying with patients who are left in conditions that are far from one's own situation (Female rater 1). Precisely such experiences may have been a significant source of the raters’ taste for the candidates’ ability to connect with and empathize with people and situations.

In summary, the MMI raters’ tastes stem from classificatory schemes related to the raters’ own incorporation and habituation of the societal structures in the Danish medical educational and health care system. The classificatory schemes concern self- and work-discipline, work-life balance, and a pragmatic approach to the complexities of medical practice.

**Enculturation of taste**

Since the raters had not been interviewing in connection with MMI tests before, they felt some uncertainty in relation to how accurate their assessments were, despite their practical and personal experiences as medical students and doctors. Several raters expressed that they developed confidence and security in their own taste during the day, among others because they gained knowledge about the assessments of their co-rater. Most raters who shared a station with a co-rater noted that they compared scores after these had been given. Generally, the co-raters experienced a great deal of consensus, especially when it came to the highest and lowest scoring candidates. These comparisons seemed to create opportunities for influencing each other and approximating - or even aligning - tastes:

Female rater 5: It was really great when we were two of us [at the station]. It seems somewhat unfair that you adapt from one situation to the next because you assess the first candidate and ask [your colleague at the station] "What was your assessment and what is mine?" and then the scale shifts. But that's probably just a condition for that kind
of tests. And if you got the chance to repeat, then you would probably do a bit better because you had tried it before and had formed an idea about how they were doing and what to look for.

In particular, the raters’ lack of previous practical experience with the MMI seemed to have an impact on how they scored the first candidates. For example, one rater explained that he had abstained from giving the highest score to his first candidate even though he immediately felt this candidate was highly capable. The rater was afraid of going through the roof of the marking scale, so he awarded the candidate the score of 6, which he later changed to seven when he realised that the candidate had delivered the best performance that day (Male rater 3). Several raters indicated that the first assessments were like fumbling in the blind. This suggests the raters lacked a sense of how each score (of the marking scale) ‘tasted’ in practice, but that this was developed concurrently with increased practical experience or that the assessment was very much relative as opposed to absolute/criteria-based in nature. Thus, the raters’ taste for ‘good’ or distaste for the ‘poor’ performances was not developed only through their own socialisation as students and doctors, but also refined from their growing practical experience as a rater in the MMI. Moreover, two raters noted how, in the beginning of the day, they felt it necessary to be strictly guided by the anchor criteria, whereas later they let themselves be guided more by their gut feeling and how they had assessed prior candidates (Male rater 3; Female rater 7). This may suggest that the raters’ taste was validated and refined as a result of the specific interactions and situations in which the raters were involved during the tests, and that the raters’ experience of making valid assessments is not only created by having access to assessment criteria, no matter how detailed they may be, but is also formed by the context-specificity of the test and the practical experiences it affords the rater. Moreover, the above may also suggest that while the taste of the raters may be a matter of instinct and immediate perception, it is not fixed/automated or sealed/inaccessible, but develops in process and is extendable provided it is subjected to attention and sensation.
In summary: Enculturation seemed to provide the raters with a collective and shared taste for the 'good' or distaste for the 'poor' applicant performances. The enculturation seems to have occurred not only during medical school and clinical work, but also to some extent during MMI interviews.

DISCUSSION

After a summary of our key findings we will compare these findings with existing literature and theory, and discuss the methodological challenges and strengths of the study. Finally, we will discuss the educational implications of this study.

Summary of key findings

This study investigated MMI raters' subjectivity with regard to their 'taste' for medical school candidates. We found that the raters spontaneously applied subjective criteria — their taste — which enabled them to assess the candidates' interpersonal attributes and to predict the candidates' future potential. The objective set-up (the anchor criteria), which structured the MMI test, may not only have helped and supported the raters to standardise their assessments; there were also indications that the anchor criteria may have limited and restricted the raters' possibilities for using the entire scope of their taste during the interviews, which resulted in uncertainly and discomfort when applying taste that did not match the objective criteria. This may have influenced the quality of their assessment. The MMI raters showed a taste for medical school candidates who displayed reflectivity, resilience, capacity to connect, empathy, alikeness and potential for being 'the good colleague'. Finally, we found that MMI raters seemed to share a collective taste for the aforementioned qualities in the candidates; hence, taste may be the result of an ongoing enculturation — long-term (the rater's age-long socialisation in medical school and in the Danish healthcare system) as well as short-term (a few hours of interaction with a co-rater in the context-specific MMI test).
Comparison of key findings with existing literature and theory

Below we compare the key findings with 1) existing literature and 2) theory on taste and subjectivity.

To the best of our knowledge, only one study in medical education (Kumar et al., 2009) has previously examined MMI raters’ experience of subjectivity during the MMI test from the point of view of the raters. Thus, we know only little about the workings of MMI raters’ subjective judgements. Our findings are consistent with those of Kumar and colleagues (2009) who found that MMI raters in their study expressed concern about the lack of opportunity to benchmark their decisions (Kumar et al., 2009). The raters in our study had similar experiences and felt an underlying uncertainty and discomfort associated with relying on subjectivity and taste.

Our study showed that MMI raters’ tastes are characterized by a preference for candidate attributes in the realm of reflectivity and empathy. Other studies found that candidates’ non-cognitive attributes such as reflectivity and empathy may be difficult to categorize and therefore difficult to fully capture with quantifiable and universal measures (Cunnington et al., 1996; Sebok et al., 2014). Kumar et al. (2009) suggested that the one-to-one interview may provide an arena for showing and assessing non-cognitive skills. Based on our findings we agree that comprehensive assessments of these attributes should be based on raters’ perceptions of the interpersonal dynamics that unfold during face-to-face encounters with candidates. This places demands on the raters’ ability to be attentive to the (often non-verbal and bodily sensed) particularities of the test situation and, as Albanese and colleagues (2003) showed, raters should add a “human touch” (Albanese et al., 2003, p. 315) to the assessment process, i.e. be sensible to and connect with the candidates at a profound interpersonal level. In line with Albanese and colleagues (2003), our findings suggested that the raters’ practical sense – and thus taste – is essential to making good assessments of candidates to medical school and that it may not be desirable to restrict raters from using their subjectivity (taste) during MMI tests. However, our results and the few studies mentioned above are only scratching the surface of the underexplored
research area of taste and subjectivity related to MMI raters' assessment of candidates to medical school. By using the sociological concept of taste as the theoretical framework in this study, we have enhanced the understanding of how rater subjectivity is played out, we next discuss our key findings with further theoretical reflections on taste and the difficulties in applying taste in cultural settings – such as medical education – where objectivity is part of the cultural forms.

Cognitive dissonance and habitus clivé

Our study endorses the theoretical assumption that taste is fundamental to any kind of human judgement and therefore cannot be bypassed by any attempt to objectify the MMI testing process. However, avoiding objectification is easier said than done. According to sociologist Nick Crossley (2013), actors (in this case MMI raters) develop positive associations to the cultural norms of alters (in this case the medical school and the Danish healthcare system). However, being positioned in the objectified set-up of the MMI test in the context of medical education and encultured in the medical world, and at the same time being told to use their ‘gut feelings’ and to take a subjective stance, may cause cognitive dissonance in the raters, because two apparently conflicting cultural norms – objectivity and subjectivity – may be difficult to apply simultaneously. Development of taste also implies distaste for conflicting cultural forms. As physicians, the raters seemed to have developed positive associations to the cultural forms in medicine: an ingrained taste for the standards of the positivist scientific belief system, an appetite for objectivity, a habit of regarding objective measurements as the highest scientific standard, and thus presumably a distaste for the claim for subjectivity in the MMI test. According to Crossley (2013), actors resolve such cognitive dissonance by trying to like conflicting cultural forms – but only in some cases do they succeed in doing so. In explaining the dynamics of taste and distaste, Crossley draws on a socio-psychological understanding of human beliefs and attitudes described in the cognitive dissonance theory by Leon Festinger. Cognitive dissonance theory emphasises beliefs as a central component of an attitude (Hogg and Vaughan, 2005). It states that cognitive dissonance is an
unpleasant state of psychological tension generated when a person has two or more cognitions that are inconsistent. Thus, people will try to reduce this tension by changing or rejecting one of the cognitions or seek consonance by derogating the source of one of the cognitions (Hogg and Vaughan, 2005). Similarly, Crossley describes how taste is both formed by social interactions and affects interpersonal encounters. Thus, taste is subject to alteration because actors tend to develop positive associations to the cultural forms that are appreciated by the social context they (try to) inhabit. From a more sociological point of view, Bourdieu (1998) argues that taste is invisible or a tacit category of perception which results from incorporated societal structures. This means that taste derives from relatively robust cognitive structures that are intimately linked to particular objective structures in the social field. Consequently, taste is not a completely fixed entity, but is shaped in interaction between habitus and field. Instead of the sociopsychological term cognitive dissonance, Bourdieu suggests the sociological notion of habitus clivé (Bourdieu, 2000) to describe a mismatch between habitus and field, that is the individual experience of a sense of self cleft due to changing conditions of existence. This may amount to a deeply felt discrepancy between the tastes (beliefs, thoughts and preferences) of two different cultural forms. Accordingly, actors (in this case: MMI raters) will try to bridge the cleft by seeking to reproduce the taste of the conflicting cultural form, while trying to be true to the inhabited cultural form. The point is that both cognitive dissonance and habitus clivé reflect the dilemma expressed by the MMI raters when they told about being uncertain about their taste and how they developed their taste during the MMI test day because they gained knowledge about the assessments of other raters and in this way sought harmony and aimed to reduce the dissonance without overriding their initial gut-feelings. In other words, the dilemmas faced by the raters are essentially underpinned by a fundamental tension between the positivist (objective) and interpretivist (subjective) world views and values regarding decision making in high stakes situations.

Our findings support the theoretical notions of cognitive dissonance and habitus clivé as important points to pay attention to when asking MMI raters – and probably also raters of
students' clinical performance in authentic settings – in medical education to rely on subjective gut feelings. This insight invites a discussion of how the MMI may make room for the raters' taste and how training of the raters may be facilitated before participation in an MMI.

Methodological strengths and challenges

The methodological strength of our study is its' foundation in a solid theoretical framework which seems to be innovative and original compared to the current research on MMI. However, our study had some challenges. Firstly, the 12 included participants in the interview study did not entirely reflect the background population (all 104 MMI raters) in terms of title and occupation. The larger representation of physicians who were residents enrolled as PhD students may be explained by this group's relatively more flexible hours and the circumstance that all interviews were scheduled to be conducted in one particular week. Conversely, the PhD students seemed to better recall the demands and conditions at medical school and the transition from medical school to clinical work as a physician. Secondly, a larger share of raters was from the domains 'resilience' and 'communication' which may have influenced the results. However, it should be mentioned that all raters were asked to score the candidates in two test domains: the primary test domain and one of the other three test domains. In this way, the raters had experience from scoring candidates in two test domains. Finally, we purposively selected twelve MMI raters who gave low scores to at least one MMI applicant and therefore we might have included the 'hawks' and not the 'doves' among the raters. If this was the case, the procedure could raise questions pertaining to generalizability, but for the purpose of this study, we sampled MMI raters who gave low as well as high scores and employed the whole rating scale. In addition, we asked the raters about both low and high score applicants. Thus, we reason that the selected participants acted as both 'hawks' and 'doves' in the MMI test. Moreover, in this qualitative study we adhered to analytical generalization (Kvale and Brinkmann, 2009) that rest upon rich contextual descriptions and theoretical analyses instead of statistical generalizability.
Educational implications: Make room for MMI raters’ taste and train the ability to taste

Dreyfus and Dreyfus’ thoughts on intuitive expertise and skilled judgement (Dreyfus and Dreyfus, 1986; Dreyfus, 2004) have been discussed in medical education as a way to understand and acknowledge how expertise develops in a process from rule-bound judgements to context-based intuitive judgements (Batalden et al., 2002; Stevens, 2002; Penã, 2010). Based on these thoughts, one may argue that the MMI raters will only be capable of making skilled judgements if they disengage themselves from rules, principles and universal explanations in the form of objective and standardised criteria, and instead make context-based intuitive judgements based on experiences from similar, prior situations. In line herein, this study shows that the raters are influenced by the complexities and particularities of their everyday practice in the form of incorporated practical experiences. When using their taste, they intuitively compare the candidate’s test performances with these practical experiences. Since tasting allows a myriad of impressions and experiences to influence one’s overall judgement, urging raters to use their taste during MMI may potentially yield more nuanced assessments of the candidates than if the assessment is primarily based on standardised objective items on a checklist. In this manner, this study substantiates the suggestion that being guided by one’s subjectivity – or taste – is not just a matter of personal bias and error, but may be a matter of practical matching, i.e. relating the holistic impressions of the candidate to the practicalities and complexities of the MMI rater’s educational and professional experiences. In this perspective, and following the example of Kumar and colleagues (2009), we question the imperative need for standardised objective criteria or even for the predetermined test domains that have hitherto characterised the MMI test format. Such criteria may restrict the raters from examining the candidates in other respects which they consider (more) important. The above arguments suggest that a more holistic measurement process or a broader unidimensional criterion for measurement such as the candidates suitability for medical school/medical practice (Sebok et al., 2014) may be expedient. This argument is supported by Cunnington et al. (1996) who suggest that global ratings are as
reliable as more traditional and detailed checklist scorings. On the other hand, one may argue that too much emphasis on the use of taste and ‘gut feelings’ may leave raters with the feeling of lacking direction and may encourage them to make snap judgements based on naïve and stereotypical assumptions (Roberts et al., 2010). This would imply that the raters’ tasting of candidates would be based on certain coagulated classification schemata of what the good candidate should be like (in a way much similar to that of the objective and standardised checklist), which risk making the raters insensible to other nuances of taste than their own. In addition, even proponents of a paradigm shift towards embracing subjective judgments in assessment recognize that the psychometric paradigm with its focus on objectivity has been important for increased fairness towards test-takers (Hodges, 2013). Consequently, an educational implication of this study would be to make (more) room for MMI raters’ taste during the test.

Studies suggest that the above-mentioned problem about raters’ naïve and stereotypical assumptions may be solved by training the raters, for instance by encouraging raters to refrain from making snap and prejudiced judgements (Robert et al., 2010), or by developing a shared understanding of what should be assessed and what could be areas of potential rater bias (Kumar et al., 2009). On the basis of our study, we suggest that rater training may also involve the refinement of the raters’ ability to taste. As suggested by Bourdieu (1998), Dreyfus and Dreyfus (1986), Polya (1958) and others, learning to become a skilled practitioner, such as a skilled MMI rater, is not only about following explicit rules or categories. Instead, it is about habituating certain classificatory schemata through practice until one becomes capable of making holistic discriminations, which, in the case of MMI raters, will mean being capable of designating candidates that match a medical career. However, whereas the habituation of classificatory schemata may make the raters become more precise, effective, quick and secure, habituation may also make them so confident in their own taste that it weakens their awareness of the particularities (and other nuances) during the test. Consequently, we suggest that effective training involves urging the raters to continue active attendance to their own tasting. This can be
further illustrated by considering the active tasting of a connoisseur (Carlisle, 2014). A connoisseur is a person who has trained tasting and is particularly competent to pass critical judgments in matters of taste. If MMI-raters train rating like a wine connoisseur, they actively develop a heightened sensitivity and ability to make judgements about their tasting. In contrast, if the raters train like the habitual drinker, they passively consume rating and their taste may coagulate into fixed classificatory schemes that are almost impossible to bypass. Using the words of Carlisle (2014, p. 83), the raters’ training “should be regarded as an elevation of habit rather than a departure from it”. Accordingly, MMI rater training may contribute to the refinement of raters’ taste and the development of connoisseurship, which involves awareness of distinctions in taste – including the distinctions between individual and shared tastes and the raters’ “principles of vision and division” (Bourdieu, 1998, p. 25). Consequently, an educational implication of this study would be to acknowledge tasting as practical expertise and to train the raters’ ability to taste.

Taste cannot (and should not) be bypassed in the assessment of students’ performance. However, rater subjectivity is still an underexplored research area and more investigations are needed in order to draw solid conclusions on the workings of raters’ taste in different educational settings, including authentic clinical workplaces.

ACKNOWLEDGEMENTS

The authors wish to extend their gratitude to the 12 MMI raters who participated in the interviews and kindly shared their experiences.

TABLES

Table 1: Demographic data for the included participants
<table>
<thead>
<tr>
<th>Participants</th>
<th>Profession</th>
<th>Academic Position</th>
<th>Overall specialty</th>
<th>MMI domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female rater 1</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Empathy</td>
</tr>
<tr>
<td>Female rater 2</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Communication</td>
</tr>
<tr>
<td>Female rater 3</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 4</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 5</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 6</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 7</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Resilience</td>
</tr>
<tr>
<td>Male rater 1</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 2</td>
<td>Biologist</td>
<td>Associate Professor</td>
<td>Biomedicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Male rater 3</td>
<td>Physician</td>
<td>Resident</td>
<td>Medicine</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 4</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 5</td>
<td>Physician</td>
<td>Resident</td>
<td>General Practice</td>
<td>Collaboration</td>
</tr>
</tbody>
</table>
### Table 2: Selection of participants

<table>
<thead>
<tr>
<th></th>
<th>All MMI raters (n=104)</th>
<th>Participants in this study (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55 (53%)</td>
<td>7 (58%)</td>
</tr>
<tr>
<td>Male</td>
<td>49 (47%)</td>
<td>5 (42%)</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician (senior)</td>
<td>20 (20%)</td>
<td>2 (16%)</td>
</tr>
<tr>
<td>Physician (PhD student)</td>
<td>44 (42%)</td>
<td>9 (75%)</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>26 (25%)</td>
<td>1 (8%)</td>
</tr>
<tr>
<td>Not known</td>
<td>14 (13%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Domains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy</td>
<td>25 (26%)</td>
<td>1 (8%)</td>
</tr>
<tr>
<td>Communication</td>
<td>23 (23%)</td>
<td>4 (33%)</td>
</tr>
<tr>
<td>Resilience</td>
<td>23 (23%)</td>
<td>6 (50%)</td>
</tr>
<tr>
<td>Collaboration</td>
<td>27 (28%)</td>
<td>1 (8%)</td>
</tr>
<tr>
<td><strong>Lowest score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>19 (19%)</td>
<td>5 (42%)</td>
</tr>
<tr>
<td>2</td>
<td>37 (37%)</td>
<td>7 (58%)</td>
</tr>
<tr>
<td>3</td>
<td>37 (37%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>4</td>
<td>6 (6%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
REFERENCES


Doi:10.1111/j.1365-2923.2009.03291.x


Table 1: Demographic data for the included participants

<table>
<thead>
<tr>
<th>Participants</th>
<th>Profession</th>
<th>Academic Position</th>
<th>Overall specialty</th>
<th>MMI domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female rater 1</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Empathy</td>
</tr>
<tr>
<td>Female rater 2</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Communication</td>
</tr>
<tr>
<td>Female rater 3</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 4</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 5</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 6</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Female rater 7</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Resilience</td>
</tr>
<tr>
<td>Male rater 1</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Medicine</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 2</td>
<td>Biologist</td>
<td>Associate Professor</td>
<td>Biomedicine</td>
<td>Resilience</td>
</tr>
<tr>
<td>Male rater 3</td>
<td>Physician</td>
<td>Resident</td>
<td>Medicine</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 4</td>
<td>Physician</td>
<td>Resident/PhD student</td>
<td>Surgery</td>
<td>Communication</td>
</tr>
<tr>
<td>Male rater 5</td>
<td>Physician</td>
<td>Resident</td>
<td>General Practice</td>
<td>Collaboration</td>
</tr>
</tbody>
</table>
### Table 2: Selection of participants

<table>
<thead>
<tr>
<th></th>
<th>All MMI raters (n=104)</th>
<th>Participants in this study (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55</td>
<td>7</td>
</tr>
<tr>
<td>Male</td>
<td>49</td>
<td>5</td>
</tr>
<tr>
<td><strong>Title</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician (senior)</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Physician (PhD student)</td>
<td>44</td>
<td>9</td>
</tr>
<tr>
<td>Associate Professor</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>Not known</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td><strong>Domains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Communication</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>Resilience</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>Collaboration</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td><strong>Lowest score</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
<td>19%</td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>37%</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>37%</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>6%</td>
</tr>
</tbody>
</table>