DUT
Dansk Universitetspædagogisk Tidsskrift

Tema Fra data til beslutninger

Årgang 14 nr. 26 / 2019

Titel **Intended and unintended test constructs in a Multiple-Mini admission Interview. A validity Study.**

Forfattere Lotte Dyhrberg O'Neill, Eva Lykkegaard, Kulamakan Kulasageram

Sidetal 66-81

Udgivet af Dansk Universitetspædagogisk Netværk, DUN

URL › http://dun-net.dk/

Intended and unintended test constructs in a Multiple-Mini admission Interview. A validity study.

Lotte Dyhrberg O'Neill[a, 1], Eva Lykkegaard[b], Kulamakan Kulasageram[c]

[a]SDU Universitetspædagogik, Syddansk Universitet, [b]Institut for Kulturvidenskaber, Syddansk Universitet, [c]The Wilson Centre for Research in Education, University of Toronto

Research article, peer-reviewed

Admission interviews in higher education may be developed with the intention to select applicants with specific personal competences not captured by traditional grade-based admission. In this study, we examined whether the data structure of multiple-mini admission interview scores supported the presence of communication, empathy, collaboration, and resilience as independent test dimensions. In addition, the associations between the interview scores and unintended test constructs (station format, pre-university grades, age, gender) were examined. Confirmatory and exploratory factor analyses and regression analyses were used to examine interview data from a cohort of Danish medical school applicants. The proposed multi-dimensionality was not supported by the data structure. The influence of the unintended constructs examined was limited or non-existing. These results are in line with the scarce existing literature. This situation makes a priori claims that the multiple-mini interview can measure multi-dimensional personal competences inadvisable, and care should be taken about what is communicated to stakeholders.

Introduction

Every year, thousands of medical school applicants around the world are selected or rejected based on their performances in admission interviews. Medical schools may aim to select and reject applicants based on specific, predefined, non-academic, personal competencies, although very little evidence exists in the literature to support that such specificity is possible (Albanese, Snow, Skochelak, Huggett, & Farrell, 2003; Eva et al., 2009; Knorr & Hissbach, 2014; Patterson et al., 2016). Leaving society and rejected applicants with the impression that specific personal competencies (such as communication skills or empathy, etc.) were assessed seems unfortunate if judgments are actually much more general in nature.

The last decades have seen the emergence of an interest in admission testing in Danish health science education (O'Neill, Christensen, Vonsild, & Wallstedt, 2014; ONeill, Vonsild, & Wallstedt, 2013; O'Neill, Hartvigsen, Wallstedt, Korsholm, & Eika, 2011; Wallstedt, 2004), and a subsequent political focus on examining the traditional admission system in Denmark (Danmarks Evalueringsinstitut, 2015). Most recently, the interest in admission testing has spread across faculties and to several large institutions in Danish higher education (Dinesen, 2018). A number of programs now use the Multiple Mini-Interview (MMI) as a selection tool, but only a

[1] Contact: ldo@sdu.dk

limited number of Danish effect studies have been published to date (Vonsild, Schibler, & Wallstedt, 2016; Danmarks Evalueringsinstitut, 2017).

The Multiple Mini-Interview (MMI) is the latest major development in selection interviews for medical education. The MMI consists of a series of independent interview stations (typically 8-12), each with different tasks to be solved, and each manned with its own rater. The test time on each station is fixed and relatively short (typically <10 minutes per station), and test takers proceed from station to station at the sound of a bell until all interview stations in the MMI have been completed. Figure 1 gives a birds-eye view of an ongoing MMI in which individual stations are separated with partitions.

Figure 1. An MMI in operation.

The MMI was originally described as being useful for testing non-academic competences believed to be important for high quality health care, such as communicative and collaborative competences etc. (Eva, Rosenfeld, Reiter, & Norman, 2004), but few published studies have examined whether such specific assumptions are warranted. There seems to be emerging evidence that MMI scores can predict subsequent pre-graduate and post-graduate performances (Knorr & Hissbach, 2014; Patterson et al., 2016; Pau et al., 2013), although it has been pointed out that presently most of this evidence arises from a single institution (Eva, Reiter, Rosenfeld, & Norman, 2004a; Eva et al., 2012; Eva et al., 2009; Knorr & Hissbach, 2014). In addition, there is evidence for stakeholder acceptability (Patterson et al., 2016; Pau et al., 2013).

When it comes to validity evidence for the 'internal structure' of MMIs (American Educational Research Association, 2014), the results are somewhat equivocal. Evidence of internal structure rooted in generalizability studies tends to display adequate generalizability when sufficient stations and raters are sampled (Knorr & Hissbach, 2014; Patterson et al., 2016; Pau et al., 2013). However, the evidence of internal structure relating to the dimensionality of MMIs remains problematic (Knorr & Hissbach, 2014; Patterson et al., 2016). Patterson et al. reviewed the literature and concluded that 'it is critically important that schools better understand what they are seeking to measure with this approach' (Patterson et al., 2016). We found only a few published studies which examine the intended multi-dimensionality of MMI scores (Hecker et al., 2009; Lemay, Lockyer, Collin, & Brownell, 2007; Oliver, Hecker, Hausdorf, & Conlon, 2014). At the same time, a few studies based on Item Response Theory (IRT) approaches indicate that MMIs perhaps measure a much broader and unidimensional con-

struct (Jones & Forister, 2011; Knorr & Hissbach, 2014; Roberts, Zoanetti, & Rothnie, 2009; Sebok, Luu, & Klinger, 2014).
Because of this general lack of global evidence for the multi-dimensionality of MMIs (Knorr & Hissbach, 2014; Patterson et al., 2016), we felt it was important to examine and publish evidence on the internal structure of MMI scores for global as well as for national and local purposes. Examining the relationships between MMI scores and potential competing constructs such as prior grades was also particularly important in a Danish context, as the test-based admission track (kvote-2 admission) is supposed to offer a real alternative to the grade-based admission track (kvote-1 admission).

Aim

The aim of this study was to examine aspects of validity in an MMI relating to the internal structure and to the relationships to other variables that could be competing constructs. The objectives were to 1) examine whether the factor structure of the MMI scores supported the presence of test domains corresponding to content themes (communication, empathy, collaboration, and resilience), and to 2) examine the association between the MMI scores and unintended competing test constructs such as the station format, age, gender and pre-university grades. The hypotheses were that i) the internal structure of the scores would reflect the four content themes as test domains, and ii) that the MMI scores were not just a reflection of demographic differences, different station formats or general academic abilities.

Method

Design

The study is an observational validation study of the scoring process in an MMI for a cohort of medical school applicants. It has both a cross-sectional element (factor analyses of MMI scores from one occasion), and a retrospective element (regression analyses of the association between competing explanatory variables and MMI scores). The theoretical framework behind the research questions examined is unified Validity Theory (Messick, 1989), where evidence of the 'internal structure' of scores and the 'relationships with other variables' count as sources of construct validity evidence (American Educational Research Association, 2014). With that framework as the backdrop, the hypotheses (i and ii above) are the validity assumptions examined (Kane, 2006).

Test takers

The intended study population was all medical school applicants and raters who participated in the MMI at Aarhus University in June 2016. MMI test takers were selected from a larger pool of approx. 1,800 medical school applicants based on best performance in a written reasoning test (the uniTEST).

Raters

The raters on the MMI stations were a mixture of biomedical lecturers, clinical teachers (practicing regional doctors), and health science doctoral students. Raters participated in a two-

hour training session which involved general information, scoring practice, reflections and discussions before participating as raters.

MMI scores

In 2016, eighty percent of medical students at Aarhus University were admitted via traditional grade-based admission, and the remaining twenty percent via test-based admission. A double-track admission system has been in place in Denmark since the mid 70's, because changing governments wanted to secure broader access to higher education with an alternative to purely grade-based admission. The test-based admission at Aarhus University Medical School consisted of three steps. Step 1 was a requirement of minimally acceptable grades from an upper secondary education. Step 2 was participation in a 95-item multiple choice test of generic reasoning and thinking across the two broad domains of mathematics/science and humanities/social sciences known as the uniTEST, which is developed by the Australian Council for Educational Research. Step 3 was participation in an MMI developed to test non-academic personal competences. It is the latter step we examine in this study.

The MMI was developed and pilot tested by the Centre for Health Sciences Education, Aarhus University in cooperation with invited stakeholders (regional clinicians, lecturers, patient organizations, the national medical association). Four relevant test themes (communication, empathy, collaboration, and resilience) were identified with the Nominal Group Technique by a Delphi group of stakeholders. Eight different MMI stations (2 per test domain) were subsequently developed and pilot tested in November 2015. Preliminary validity evidence has been published relating to content, internal structure, relations with other variables, and consequences of testing based on the pilot test results (Andreassen et al., 2016). A generalizability study (internal structure validity evidence) revealed acceptably high generalizability coefficients (G=0.85 using two raters per station). However, if test results were to be generalized to any other set of MMI stations with alternative station content (station as a random facet), the MMI would have had to consist of at least 25 stations in order to reach a G-coefficient of 0.75 when using two raters per station. In addition, a disattenuated correlation analysis showed only low-moderate correlation coefficients between station pairs with the same theme but different content. In some instances, stations with different themes but similar station format (e.g. 'the situational judgment' format) displayed even stronger correlations (Andreassen et al., 2016; Eva & Macala, 2014), which could suggest a potential station format effect. Examination of the pilot MMI scores' relations to other variables found no significant correlations with selected competing constructs, such as prior academic competences (pre-university grade point averages), gender, or age (Andreassen et al., 2016). However, the pilot test results were based on a small sample (n=26) of simulated applicants (volunteering junior medical students) and not necessarily representative of the broader applicant pool. After minor adjustments, the MMI was re-administered in June 2016, this time to real medical school applicants participating in the test-based ('kvote-2') admission track at Aarhus University medical school. The development and pilot testing of the stations has been described in a published report (Andreassen et al., 2016). Table 1 gives an overview of the scales scored and the station format used on each of the eight stations. In addition to the four themes (intended test domains), raters were also asked to rate perceived participant suitability for medicine. This scale gave raters the opportunity to convert important (positive or negative) observations outside the realm of the station theme to an independent score. Raters were explained that they

were allowed to use their subjectivity on the suitability scale, and that the scores on this scale were allowed to deviate from the scores given on the other two scales.

Table 1. Distribution of scales scored and station formats in the MMI stations.

Station	Scales	Station format
1	Resilience, communication, suitability for medicine	Behavioral Interview
2	Resilience, communication, suitability for medicine	Situational Judgment
3	Communication, empathy, suitability for medicine	Situational Judgment
4	Communication, empathy, suitability for medicine	Practical/Social task
5	Empathy, communication, suitability for medicine	Situational Judgment
6	Empathy, communication, suitability for medicine	Situational Judgment
7	Collaboration, communication, suitability for medicine	Practical/Social task
8	Collaboration, communication, suitability for medicine	Practical/Social task

Note: The first scale listed for each station designates the theme for which station content was developed.

In the station format 'Behavioral Interview' (table 1), participants were asked to reflect on a previous personal experience of the station theme. In the 'Situational Judgment' station format, on the other hand, participants were asked to discuss with the interviewer either a written or a video case reflecting a relevant challenging situation. The 'Practical/Social task' station format involved solving a given practical problem in collaboration with an actor or a fellow participant. On each station, participants received a score on three scales, i.e., in 2 themes as well as a score for general suitability for medicine. This yielded 24 individual scores or data points per participant in total. Each data point was scored on a 7-point Likert scale ranging from 1 (Completely unsatisfactory) to 7 (Excellent). The scores were recorded on a paper scoresheet by raters during the MMI and subsequently transferred to an electronic admission database, and the electronic recordings were double-checked for errors. A generalizability study of the pilot test data showed a phi coefficient of 0.74 with one rater per station, and 0.85 using two raters per station (Andreassen et al., 2016). On the real admission test day, the stations were therefore manned with either one or two raters. For admission process purposes, a station score was calculated for each participant as the average of all scores given on a station across scales and raters (i.e. a number between 1 and 7). The total MMI score for a participant was the sum of the 8 individual station scores (i.e. a number between 8 and 56).

Demographic variables

Applicants were registered in the admission database with their unique Personal Identification Numbers (PIN), which all citizens have. Birthday/age and gender variables were extracted from this 10-digit PIN number. Applicants' pre-university grade-point averages were also registered in the admission database and extracted for the purpose of this study. Examining age and gender as important competing and unintended test constructs is relevant, because Danish universities are not allowed to discriminate applicants on age and gender in admissions. Similarly, if applicants' upper-secondary school GPAs were found to be an influential predictor of MMI performance, it would be problematic not just from a validity perspective as an unintended competing construct, but also for the practical purpose of having a double-track admission system in operation.

Data collection

All data was collected by the administrative leader of the admission process. The researchers received all data from the administrative leader of admissions in June 2016.

Ethics

The project was exempt from ethics review by the regional (biomedical) ethics committee as database studies and quality assurance studies do not require their permission. We obtained permission to conduct the research study from the internal legal review board at Aarhus University. The researchers adhered to the ordinary rules in the Personal Data Act regarding appropriate data handling in research projects.

Analyses

To determine whether factor analysis was at all appropriate for the data, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was examined. If the KMO was >0.5 a Confirmatory Factor Analyses (CFA) would be used to test whether the intended test domains (Model A in figure 2) or the station formats (Model B in figure 2) best explained the data observed (Field, 2009).

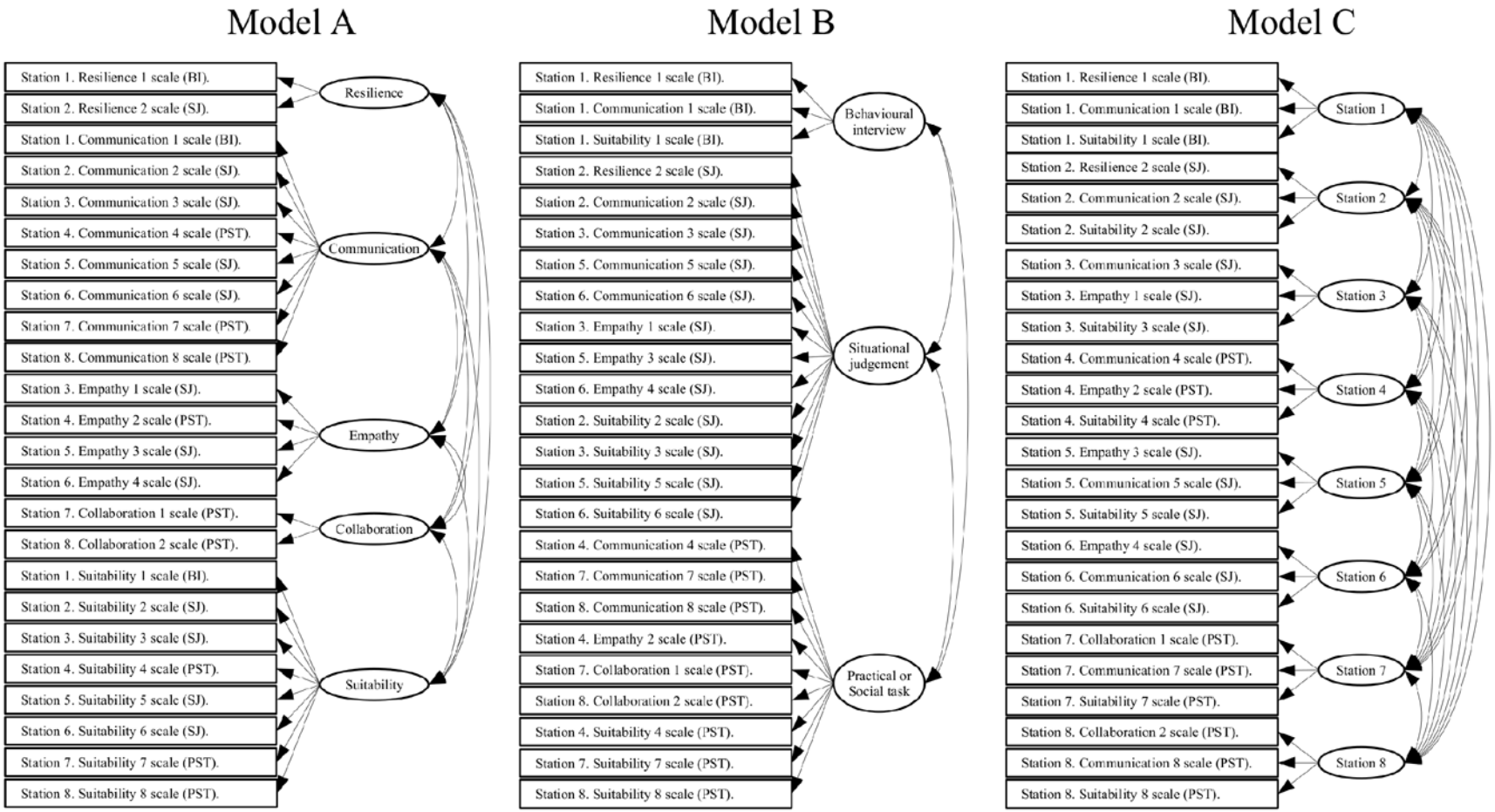

Figure 2. Models examined in the factor analyses.

If neither model was found suitable, it was decided that Exploratory Factor Analysis (EFA) was to be conducted in order to propose a better model. Exploratory factor analysis was conducted in SPSS (IBM, 2012) on the set of 24 items. A promax oblique rotation was chosen and the number of factors was decided quantitatively based on Kaiser's criterion (eigenvalues above 1) and qualitatively based on the interpretability of the factors, meaning whether each factor could be sensibly described and named separately from each other (Field, 2009). Internal consistencies of factors were examined with Cronbach's alpha. The suitability of all models examined was tested by CFA with SPSS-AMOS (Arbuckle, 2012).
Linear regression analyses were used to examine the association between potential competing constructs (age, gender, and GPA) and the MMI scores. All univariate predictors with $p<0.10$ were included in a multivariate explanatory model. We used IC STATA 14 for regression analyses and model checking procedures.

Results

175 applicants (93 males and 82 females) participated in the MMI on the test day, and their characteristics in terms of the other variables examined are summarized in table 2.

Table 2. MMI participants' characteristics (n=175).

	n	Median	5% percentile	95% percentile	Min-max
Age (years)	175	21.34	19.33	28.52	18.68-49.20
Pu-GPA	170[a]	9.20	6.70	10.90	6.00-12.40
MMI scores	175	39.33	29.33	46.83	21.50-50.00

Note: n= number of observations, SD= standard deviation, Pu-GPA=pre-university grade point average, MMI=multiple mini interview.
[a]Five MMI participants were foreign applicants without a Danish pu-GPA.

Both models A and B resulted in poor model fit measures (see Table 3), and model respecifications of both models allowing for co-variations between items did not help sufficiently.

Table 3. Model Fit measures for models A, B, and C.

	CMIN		RMR.GFI		Parsimony-Adjusted Measures	RMSEA	
	p	CMIN/DF	GFI	AGFI	PCFI	RMSEA	PCLOSE
Model A	<0.001	15.49	0.42	0.28	0.20	0.21	<0.001
Model B	<0.001	12.03	0.45	0.34	0.36	0.25	<0.001
Model C	<0.001	1.41	0.87	0.83	0.80	0.05	0.549

Note: CMIN = The Chi-square, CMIN/DF = the minimum discrepancy, RMR = the root mean square residual, GFI = the goodness of fit, AGFI = the adjusted goodness of fit, PCFI = the parsimonious comparative fit index, RMSEA = the root mean square error of approximation and PCLOSE is a test for the null hypothesis for RMSEA. The selected threshold values were: p-value > 0.05, CMIN/DF < 5.00, GFI ≥ 0.80, AGFI ≥ 0.80, PCFI ≥ 0.80. RMSEA < 0.08. PCLOSE > 0.05.

Therefore, an exploratory factor analysis was conducted. This resulted in a convincing eight-factor structure (model C in figure 2) accounting for approximately 92% of the total variance and with a high interpretability, as the eight factors each represented exactly one MMI station's three domain scores. The eight factors only correlated weakly (<0.45) with each other. In addition, the internal consistencies of each of the eight factors were all excellent (above 0.9). The quality of the proposed model C (figure 2) was subsequently tested using a confirmatory factor analysis. All threshold values except the p-value indicated good model fit for model C (table 3). However, since the p-value is rather sensitive to sample size, it is defendable to disregard it in a situation like ours, where there is otherwise evidence of good model fit.

Prior grades did not predict MMI performance, whereas a combination of gender and biological maturity was able to explain an estimated 12.6% of the variance in overall MMI scores (table 4).

Table 4. Background characteristics as predictors of MMI performance.

		Univariate analyses			Final model		
Predictor	n	ß [$CI_{95\%}$]	p	R^2	ß [$CI_{95\%}$]	p	R^2
Gender (female)	175	1.60 [0.05-3.14]	0.043	0.023	2.04[0.53-3.54]	0.008	0.126
Age (years)	175	0.42 [0.24-0.59]	<0.001	0.089	0.45[0.28-0.63]	0.000	
Pu-GPA	170	0.28 [-0.30-0.85]	0.347	0.005	-	-	-

Note: ß =regression coefficient, $CI_{95\%}$=95% confidence interval, R^2=the coefficient of determination.

As seen in the final model (table 4), a female participant in the MMI tended to score 2.04 points higher on average than a male participant of the same age (table 4), corresponding to a difference of <5% of the possible range (7-56 points) of applicants' MMI scores. Likewise, an age difference of 1 year between two participants of the same gender tended to result in an MMI score of 0.45 in favor of the older participant.

Discussion

This study aimed to examine aspects of validity in an MMI relating to the internal structure and relationships with other variables which could be competing constructs. Our results show that MMI performance appeared to be overwhelmingly station-specific. The proposed multi-dimensionality was not supported by the data structure. Neither station format nor pre-university grades appeared to be competing constructs and the influence of gender and age on test scores was limited.

Dimensionality

The test domains examined in this MMI have also been in use elsewhere (Cleland, Dowell, McLachlan, Nicholson, & Patterson, 2012; Dowell, Lynch, Till, Kumwenda, & Husbands, 2012; K. W. Eva et al., 2004; Gafni, Moshinsky, Eisenberg, Zeigler, & Ziv, 2012; Harris & Owen, 2007; Lemay et al., 2007), but little evidence in support of their actual existence has been published so far. The question of dimensionality of MMIs therefore remains largely unresolved and the existing evidence is sparse, and results are equivocal (Hecker et al., 2009; Lemay et al., 2007; Oliver et al., 2014). The results of this study did not support the presence of four test domains (communication, empathy, collaboration, and resilience) corresponding to the themes used in the blueprint for the development of station content (Andreassen et al., 2016). On the contrary, we found very high correlations (>0.90) between different subscales used on the same stations, which indicates raters' problems with differentiating between multiple test dimensions in a situation. The results generally fit well with what is known about context and context specificity (*situation* or *state*) as the dominant influence on human behavior (Ross & Nisbett, 1991). Findings in social and cognitive psychology have suggested that competences such as problem solving, professionalism, communication, team performance, etc. are highly

context-dependent rather than generic (Eva, 2003; C. P. M. van der Vleuten, 2014). This does not necessarily mean that these personal qualities and the traits/attributes cannot be measured in meaningful ways. For example, models of personality theory have recently begun to engage with the relevance of context in shaping behaviors that are affected by behavioral traits, which are more akin to 'mean norms' with a variability affected by contextual and other interacting factors (Ferguson & Lievens, 2017). Detecting 'stable measurable traits' in admissions tools such as MMIs must address this complex interplay between contextual factors and underlying traits to truly establish construct validity.

This complexity is also reflected in the heterogeneous and conflicting conclusions in previous research on MMI construct validity (Hecker et al., 2009; Lemay et al., 2007; Oliver et al., 2014). Lemay and colleagues evaluated a 10-station MMI which was assumed to test different domains on each station. Their exploratory factor analysis confirmed a 10-factor solution which supported the proposed multi-dimensionality (Lemay et al., 2007). However, the 10 dimensions examined were nested in stations, and so could not be disentangled from situation or context. A second study examined the factor structure of a 5-station MMI used in admission to veterinary school (Hecker et al., 2009). While the intention was to examine the intended five non-academic domains in the MMI, the factor analysis revealed only two factors. Domains such as empathy and moral/ethical reasoning were not discrete and independent as initially assumed (Hecker et al., 2009). In the third study, researchers examined MMI scores for the presence of two conceptually distinct a priori identified, non-academic attributes to be tested. While the data was best explained by a two-factor model rather than a one-factor model, the two proposed domains ('Oral Communication' and 'Problem Evaluation' respectively) correlated very highly (0.87). According to the authors themselves, this limited the ability to conclude that two independent factors were assessed (Oliver et al., 2014). At this point, therefore, there seems to be conflicting evidence from factor analytic studies (including ours) as to the ability of the evaluated MMIs to test multiple dimensions. Similarly, evaluations of the factor structures of Objective Structured Clinical Examinations (OSCE) - the station-based assessment format which inspired the development of the MMI - have also shown heterogeneity with some aligning well with broad factors underlying multiple stations, while others reflect station-specific factors (Volkan, Simon, Baker, & Todres, 2004).

The equivocal results from the published factor analytic studies are further buttressed by a smaller number of studies based on multi-facet Rasch and item-response theory approaches, which have suggested that some MMIs measure a much broader and unidimensional construct (Knorr & Hissbach, 2014). These studies have suggested unidimensional constructs entitled 'entry-level reasoning skills in professionalism', 'latent professional potential', 'suitability for medical school'/'professionalism' (Jones & Forister, 2011; Roberts et al., 2009; Sebok et al., 2014). In some contexts, 'general suitability' scales have also been used purposely (Dore et al., 2010; Eva, Reiter, Rosenfeld, & Norman, 2004b; Knorr & Hissbach, 2014). In its basic form, an MMI can be viewed as a serial oral examination, a test format which has previously been shown to be influenced by examinees 'verbal style', their 'capacity to formulate ideas' and their 'communication skills' (Davis & Karunathilake, 2005). It is perhaps also possible that oral communication could be a quite influential underlying unidimensional construct being measured in MMIs. The study by Oliver et al. (2014) described above would certainly seem to indicate difficulties in disentangling oral communication from another test construct (problem evaluation).

Competing constructs

We also examined the potential influence of unintended competing test constructs. Neither station format nor pre-university grades appeared to predict MMI scores, and the influence of sex and age on test scores was limited (table 4). Finding no/weak correlations between these variables and the MMI scores is acceptable, because it indicates that both confounding and sex discrimination were probably not serious validity issues. Others have also found the correlation between MMI scores and prior grades to be insignificant (Eva et al., 2012). The limited influence of interviewees' gender has also been shown in other contexts (Eva et al., 2012; K. W. Eva et al., 2004). But overall, there appears to be relatively little focus on examining the influence of potential confounders and bias in the literature (Eva et al., 2012; K. W. Eva et al., 2004; Moreau, Reiter, & Eva, 2006), which makes it difficult for us to compare our results with the literature to any greater extent. Examination of potential confounders is important both in future validity research (American Educational Research Association, 2014; Kane, 2006; Knorr & Hissbach, 2014), and for ethical and legal purposes in local contexts.

Future research

Further work is necessary to understand the conditions under which MMIs relate to a few underlying factors and subsequent constructs, versus when they represent high levels of multi-dimensionality related to stations. Moreover, underlying relationships between stations and other theoretical sources of validity (e.g., modern personality theory) may be worth considering. Hidden in the debate on measurement is also the role of raters. Rater judgement and decision-making in assessment is receiving renewed attention in medical education (Gauthier, St-Onge, & Tavares, 2016; Tavares, Ginsburg, & Eva, 2016). Raters' idiosyncrasies, cognitive limitations, and cultural attitudes play a role in informing judgement above and sometimes far beyond the raters' abilities to judge candidate performance (Sebok & Syer, 2015; Gauthier et al., 2016). A recently published qualitative study of 12 of the raters in the MMI examined in our study found that raters spontaneously applied subjectivity criteria (their 'taste') to the assessment of applicants. They seemed to share a taste for certain qualities in the candidates (e.g., reflectivity, resilience, empathy, contact, alikeness, 'the good colleague') (Christensen, Lykkegaard, Lund, & O'Neill, 2017). Which roles rater idiosyncrasy, subjectivity, and decision-making have in affecting the measured dimensions of MMIs is still an open question worth further exploration. Raters typically infuse assessments with both error variance (bias) and with necessary quality (their expertise) simultaneously. This is a Gordian knot which has yet to be solved in any assessment of complex competences (Govaerts & Vleuten, 2013; Hodges, 2013; C. P. van der Vleuten et al., 2012; C. P. M. van der Vleuten, 2014).

Implications for practice

On a practical note, based on our evaluation of the current literature as discussed above and our own results, as well as current thoughts on validity (American Educational Research Association, 2014; Kane, 2006), programs should probably be quite cautious about claiming to test specific traits without local evidence in support of such claims. The challenge is to remember to distinguish between station themes and content used and theoretical constructs, and to be careful about the communication with stakeholders – particularly the rejected applicants - when explaining what MMI results and decisions mean and do not mean. In the

meantime, the purpose of the MMI should probably be described in much broader terms, for example, along these lines: The MMI aims to collect information concerning personal qualities in the broadest sense, and across a range of different situations perceived to be of relevance for the culture and the society in which admitted applicants will be practicing.

Limitations

Measurement will necessarily depend heavily not only on station content and situation, but also on many other aspects such as the rater population and the local medical education culture which creates value judgements around appropriate and inappropriate displays of personal qualities. This may well affect the transferability of our conclusions to other contexts. This study may also be limited by its use of a smaller sample of test takers representing only one admission cohort in one medical school. Finally, it was also a limitation that only smaller sample sizes of stations assessed each of the four domains examined, although this shortcoming is probably quite representative of the situation in many contexts.

Conclusion

This study adds to the growing body of literature on the assessment of personal qualities in admissions and specifically on the construct validity evidence for the use of MMIs. Our evaluation of the MMI at Aarhus University medical school suggests that factors such as age and gender only contributed to MMI score variance to lesser extents, and the dimensions measured in our MMI were likely highly context-specific manifestations of personal qualities within each MMI station. This result - and existing international evidence on MMI dimensionality - strongly suggests *against* an unquestioning acceptance of claims of stable multi-dimensional personal qualities being assessed in MMIs. Local evaluations of MMIs should carefully consider the contextual elements of implementation, the communication about what is being measured, and empirical verification of measurements.

Declaration of interest

The authors declare no conflict of interest.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, 78(3), 313-321.

American Educational Research Association, A. P. A., National Council of Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington (DC): American Educational Research Association.

Andreassen, P., Pedersen, K., Jensen, R. D., Møller, J. E., Carlsen, C. G., & O'Neill, L. (2016). Optagelsessamtaler på medicin-studiet ved Aarhus Universitet.

Arbuckle, J. L. (2012). *IBM® SPSS®AMOS 21. User's guide*. Chicago: IBM.

Christensen, M. K., Lykkegaard, E., Lund, O., & O'Neill, L. D. (2018). Qualitative analysis of MMI raters' scorings of medical school candidates: A matter of taste? *Advances in Health Sciences Education*, 23(2), 289-310.

Cleland, J., Dowell, J., McLachlan, J., Nicholson, S., & Patterson, F. (2012). Research report: Identifying best practice in the selection of medical students (literature review and interview survey). *London: General Medical Council*.

Danmarks Evalueringsinstitut. (2015). *Universiteternes organisering af optag til bacheloruddannelserne*. København: Danmarks Evalueringsinstitut.

Danmarks Evalueringsinstitut. (2017). *Effekten af optagelsessamtaler på læreruddannelsen*. København: Danmarks Evalueringsinstitut.

Davis, M. H., & Karunathilake, I. (2005). The place of the oral examination in today's assessment systems. *Medical Teacher*, 27(4), 294-297.

Dinesen, T. (2018). Testbaseret optag vinder frem. *Magisterbladet*, 3, 7.

Dore, K. L., Kreuger, S., Ladhani, M., Rolfson, D., Kurtz, D., Kulasegaram, K., Cullimore, A. J., Norman, G. R., Eva, K. W, Bates, S., .Reiter, H. I. (2010). The Reliability and Acceptability of the Multiple Mini-Interview as a Selection Instrument for Postgraduate Admissions. *Academic Medicine,* 85(10), S60-S63. doi:10.1097/ACM.0b013e3181ed442b

Dowell, J., Lynch, B., Till, H., Kumwenda, B., & Husbands, A. (2012). The multiple mini-interview in the UK context: 3 years of experience at Dundee. *Medical Teacher*, 34(4), 297-304. doi:10.3109/0142159X.2012.652706

Eva, K. W. (2003). On the generality of specificity. *Medical Education*, 37(7), 587-588. doi:10.1046/j.1365-2923.2003.01563.x

Eva, K. W., & Macala, C. (2014). Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. *Medical Education*, 48(6), 604-613. doi:10.1111/medu.12402

Eva, K. W., Reiter, H. I., Rosenfeld, J., & Norman, G. R. (2004a). The Ability of the Multiple Mini-Interview to Predict Preclerkship Performance in Medical School. *Academic Medicine*, 79(10), S40-S42.

Eva, K. W., Reiter, H. I., Rosenfeld, J., & Norman, G. R. (2004b). The Relationship between Interviewers' Characteristics and Ratings Assigned during a Multiple Mini-Interview. *Academic Medicine*, 79(6), 602-609.

Eva, K. W., Reiter, H. I., Rosenfeld, J., Trinh, K., Wood, T. J., & Norman, G. R. (2012). Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. *JAMA*, 308(21), 2233-2240. doi:10.1001/jama.2012.36914

Eva, K. W., Reiter, H. I., Trinh, K., Wasi, P., Rosenfeld, J., & Norman, G. R. (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education*, 43(8), 767-775. doi:10.1111/j.1365-2923.2009.03407.x

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004). An admissions OSCE: the multiple mini-interview. *Med Educ*, 38(3), 314-326.

Ferguson, E., & Lievens, F. (2017). Future directions in personality, occupational and medical selection: myths, misunderstandings, measurement, and suggestions. *Advances in Health Sciences Education*, 22(2), 387-399. doi:10.1007/s10459-016-9751-0

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D., & Ziv, A. (2012). Reliability estimates: behavioural stations and questionnaires in medical school admissions. *Medical Education*, 46(3), 277-288.

Gauthier, G., St-Onge, C., & Tavares, W. (2016). Rater cognition: review and integration of research findings. *Medical Education*, 50(5), 511-522. doi:10.1111/medu.12973

Govaerts, M., & Vleuten, C. P. (2013). Validity in work-based assessment: expanding our horizons. *Medical Education*, 47(12), 1164-1174.

Harris, S., & Owen, C. (2007). Discerning quality: using the multiple mini-interview in student selection for the Australian National University Medical School. *Medical Education*, 41(3), 234-241.

Hecker, K., Donnon, T., Fuentealba, C., Hall, D., Illanes, O., Morck, D. W., & Muelling, C. (2009). Assessment of applicants to the veterinary curriculum using a multiple mini-interview method. *Journal of Veterinary Medical Education*, 36(2), 166-173.

Hodges, B. (2013). Assessment in the post-psychometric era: Learning to love the subjective and collective. *Medical Teacher*, 35(7), 564-568. doi:10.3109/0142159X.2013.789134

Jones, P. E., & Forister, J. G. (2011). A comparison of behavioral and multiple mini-interview formats in physician assistant program admissions. *The Journal of Physician Assistant Education*, 22(1), 36-40.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Westport: ACE/Praeger.

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. *Medical Education*, 48(12), 1157-1175. doi:10.1111/medu.12535

Lemay, J. F., Lockyer, J. M., Collin, V. T., & Brownell, A. K. W. (2007). Assessment of non-cognitive traits through the admissions multiple mini-interview. *Medical Education*, 41(6), 573-579.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Moreau, K., Reiter, H., & Eva, K. W. (2006). Comparison of aboriginal and nonaboriginal applicants for admissions on the Multiple Mini-Interview using aboriginal and nonaboriginal interviewers. *Teaching and Learning in Medicine*, 18(1), 58-61.

Oliver, T., Hecker, K., Hausdorf, P. A., & Conlon, P. (2014). Validating MMI scores: are we measuring multiple attributes? *Advances in Health Sciences Education*, 19(3), 379-392.

O'Neill, L. D., Christensen, M. K., Vonsild, M. C., & Wallstedt, B. (2014). Program specific admission testing and dropout for sports science students: a prospective cohort study. Dansk Universitetspædagogisk Tidsskrift, 9(17), 55-70.

O'Neill, L., Hartvigsen, J., Wallstedt, B., Korsholm, L., & Eika, B. (2011). Medical school dropout-testing at admission versus selection by highest grades as predictors. *Medical Education*, 45(11), 1111-1120.

ONeill, L. D., Vonsild, M. C., & Wallstedt, B. (2013). Kvote 2 optagelse og akademiske præstationer: Hvor stor betydning har det adgangsgivende eksamenssnit? Dansk Universitetspædagogisk Tidsskrift, 8(14), 86-99.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, 50(1), 36-60. doi:10.1111/medu.12817

Pau, A., Jeevaratnam, K., Chen, Y. S., Fall, A. A., Khoo, C., & Nadarajah, V. D. (2013). The Multiple Mini-Interview (MMI) for student selection in health professions training – A systematic review. *Medical Teacher*, 35(12), 1027-1041. doi:10.3109/0142159X.2013.829912

Roberts, C., Zoanetti, N., & Rothnie, I. (2009). Validating a multiple mini-interview question bank assessing entry-level reasoning skills in candidates for graduate-entry medicine and dentistry programmes. *Medical Education*, 43(4), 350-359.

Ross, L., & Nisbett, R. (1991). The person and the situation: Perspectives of Social PsychologyMcGraw–Hill. *New York*.

Sebok, S. S., Luu, K., & Klinger, D. A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. *Advances in Health Sciences Education*, 19(1), 71-84.

Sebok, S. S., & Syer, M. D. (2015). Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. *Academic Medicine*, 90(11), S50-S55.

Tavares, W., Ginsburg, S., & Eva, K. W. (2016). Selecting and simplifying: Rater performance and behavior when considering multiple competencies. *Teaching and Learning in Medicine*, 28(1), 41-51.

van der Vleuten, C. P. M., Schuwirth, L., Driessen, E., Dijkstra, J., Tigelaar, D., Baartman, L., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205-214.

van der Vleuten, C. P. M. (2014). When I say … context specificity. *Medical Education*, 48(3), 234-235. doi:10.1111/medu.12263

Volkan, K., Simon, S. R., Baker, H., & Todres, I. D. (2004). Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. *Advances in Health Sciences Education*, 9(2), 83-92.

Vonsild, M. C., Schibler, A. H., & Wallstedt, B. (2016). Testbaseret optag til videregående uddannelser. *Dansk Universitetspaedagogisk Tidsskrift*, 11(20), 130-143.

Wallstedt, B. (2004). Optagelse af studerende til lægeuddannelsen. *Ugeskrift for læger,* 166, 1980-1983.