European Journal of Teacher Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cete20

Exploring the reliability and validity of multiple mini-interviews in admission to teacher education

Henna Vilppu, Eero Laakkonen, Eeva Haataja, Asko Tolvanen & Riitta-Leena Metsäpelto

To cite this article: Henna Vilppu, Eero Laakkonen, Eeva Haataja, Asko Tolvanen & Riitta-Leena Metsäpelto (04 May 2024): Exploring the reliability and validity of multiple mini-interviews in admission to teacher education, European Journal of Teacher Education, DOI: 10.1080/02619768.2024.2351068

To link to this article: https://doi.org/10.1080/02619768.2024.2351068

Published online: 04 May 2024.

EUROPEAN JOURNAL OF TEACHER EDUCATION
https://doi.org/10.1080/02619768.2024.2351068

Routledge
Taylor & Francis Group

Exploring the reliability and validity of multiple mini-interviews in admission to teacher education

Henna Vilppu[a], Eero Laakkonen[a], Eeva Haataja[b], Asko Tolvanen[c] and Riitta-Leena Metsäpelto[d]

[a]Department of Teacher Education and Centre for Research on Learning and Instruction (CERLI), University of Turku, Turku, Finland; [b]Faculty of Educational Sciences, University of Helsinki, Helsinki, Finland; [c]Methodology Centre for Human Sciences, University of Jyväskylä, Jyväskylä, Finland; [d]Department of Teacher Education, University of Jyväskylä, Jyväskylä, Finland

ABSTRACT

This study addressed the need to explore the reliability and validity of the recently renewed aptitude test based on the multiple mini-interview (MMI) format in Finnish teacher education programmes. Utilising a nationwide dataset ($N = 3306$), we explored the reliability of the format using intraclass correlations and multilevel modelling. Furthermore, we validated the intended structure of the MMI with confirmatory factor analysis and contrasted the scores of the MMI with other admission scores, i.e. cognitive measures. The results showed mostly small effects of clustering of the applicants according to the different interviewers and circuits. The largest variance component for the MMI total score was the applicant (73.3%), followed by measurement error (15.09%). The intended structure was validated, and the MMI seemed to measure different attributes from other admission measures. Thus, the study demonstrated satisfactory reliability and validity of MMI in admission to teacher education, yet the stations need further development.

ARTICLE HISTORY
Received 12 June 2023
Accepted 28 April 2024

KEYWORDS
Multiple mini-interviews; student admission; initial teacher education

Introduction

In Finland, many teacher education programmes have long maintained their place as the most competitive university programmes, with acceptance rates of less than 10%. This makes the admission process a high-stakes situation (Rees et al. 2016) and places special importance on the selection: how to select the most suitable candidates for teacher education from among thousands of applicants (Lehane, Lysaght, and O'Leary 2023). Recently, admission to teacher education programmes in Finland has been significantly developed and unified among the eight universities offering initial teacher education (ITE). Consequently, university-specific aptitude tests and personal or group interviews have been replaced by a common aptitude test based on the multiple mini-interview (MMI) format, by which the applicant can apply to almost any of the teacher education programmes.

CONTACT Henna Vilppu henna.vilppu@utu.fi Department of Teacher Education & Centre for Research on Learning and Instruction (CERLI), University Research Fellow, University of Turku, Turku, Finland

The basic idea of the MMI method is that the applicant goes through several stations with different interviewers, and multiple attributes are assessed using a predefined, structured scoring rubric (Eva et al. 2004). These multiple interviews should reduce the effect of chance and interviewer or situational biases; this is seen as an important asset of MMI compared to traditional personal semistructured or unstructured interviews.

MMI has been widely used and studied in the context of medical education, but relatively little is known about its application in admission to teacher education. However, a couple of recent studies have shown promising results concerning its reliability and interviewer and applicant perceptions of the procedure (Metsäpelto, Utriainen, et al. 2022), as well as its ability to measure actual dispositions in contrast to work experience (Salingré and MacMath 2021). Nevertheless, the utility of MMI in teacher education admissions calls for more research. The purpose of this study was to fill this gap and shed light on the reliability and validity of the MMI method as one part of the admission process with an extensive sample, the entire national population of applicants in Finland. Specifically, we aimed to study whether MMI demonstrates satisfactory reliability, whether it can measure multiple noncognitive attributes, and whether it assesses different attributes from the other admission measures.

The MMI as a method in student selection

Multiple approaches have been used in student selection. In the related literature, a distinction is typically made between measuring the cognitive and noncognitive attributes of applicants (e.g. Bardach, Klassen, and Perry 2022; Reiter et al. 2007). Cognitive attributes usually refer to intellectual prowess, measured by high school grade point average (GPA), or performance on a certain knowledge or cognitive processing test, while noncognitive attributes are often used to encapsulate all the other desired qualities in an applicant (Eva et al. 2004). Previous studies on student admission to ITE do not show very promising results for selection solely based on cognitive attributes (e.g. Bardach and Klassen 2020; D'Agostino and Powers 2009) but encourage the use of broader approaches that include specific noncognitive measures, such as psychological teacher characteristics (Bardach, Klassen, and Perry 2022). However, identifying suitable tools for the evaluation of noncognitive attributes seems to be challenging (Klassen et al. 2020).

In the domain of medical education, the MMI format was developed to tackle problems related to measuring noncognitive attributes on a single measurement occasion, such as personal or group interviews (Eva et al. 2004; Reiter et al. 2007). The MMI is a highly structured selection method in which applicants rotate through a series of stations where they respond to predefined questions on a topic, problem, or case-based scenario. Furthermore, interviewers may assess candidates on multiple attributes, both within and across interview stations (Oliver et al. 2014).

Structure and design of the MMI

The MMI utilises multiple stations with different interviewers who independently score each applicant on a standard rating scheme. Thus, it applies a multiple independent sampling methodology (Hanson et al. 2012). There has been slight variation in the number and duration of the stations, which often represents a compromise between

feasibility (e.g. resources, participant fatigue) and reliability (Dodson et al. 2009). Typically, medical schools have adopted MMIs with six to twelve stations, with durations of 5 to 15 minutes (Knorr and Hissbach 2014). Concerning reliability, a station time of 5 to 6 minutes has been proven sufficient (e.g. Dodson et al. 2009; Knorr and Hissbach 2014).

The stations are designed to measure, for example, the applicant's ability to solve problems while communicating their ideas clearly, as opposed to specific learned knowledge (Eva et al. 2004). Pau et al. (2013) provided evidence that the MMI can assess critical thinking ability, communication skills, professionalism, and attitudes towards certain ethical and social dilemmas. Thus, the MMI aims to measure multiple noncognitive attributes, which are believed to be important for and predictive of future performance in the profession (Oliver et al. 2014). However, few studies have reported identifying the target attributes through literature research and stakeholder analysis (Knorr and Hissbach 2014).

The MMI stations typically require applicants to respond to a prompt, which might be a picture, an article, or a scenario (Salingré and McMath 2021). Stations may utilise different task formats, such as discussing a topic or dilemma with the interviewer, answering standardised questions, collaborating with other applicants, or answering problem-solving tasks (Knorr and Hissbach 2014). The applicant's performance on stations is typically rated on anchored Likert scales. Station scores may be measured either on a single scale or by summing or aggregating several subscales.

The MMI is a resource-intensive method of assessment as it requires extensive human resources and space, which may affect, for example, the number of stations arranged (Uijtdehaage, Doyle, and Parker 2011; Yamada et al. 2017). Setting up an adequate number of stations and questions within stations, as well as providing a sufficient amount of training for the examiners, can improve the reliability of the MMIs but may be too costly (Roberts et al. 2008). To decrease the financial costs of the MMI, Rosenfeld et al. (2008) suggested recruiting existing faculty to be examiners, using university premises, and reducing the duration of stations rather than their number.

Reliability and validity of MMIs

In a high-stakes situation, issues of reliability and validity require special attention. Generally, reliability refers to the consistency of the measures used, while validity refers to whether the items used measure what they are intended to measure. Validity may include a number of more specific components, such as content validity (consistency between the station scenario and the target attribute to be measured) and construct validity (consistency between the scores and the target attribute; Hecker et al. 2009). In addition, the MMI literature often discusses predictive, discriminant, and face validity. Predictive validity determines the extent to which MMI identifies applicants who will display the desired attributes during training and throughout their professional practice (Cameron et al. 2017). Discriminant validity contrasts MMI with other selection measures (Breil et al. 2020), and face validity refers to the acceptability of the MMI by interviewers and applicants (Hecker et al. 2009).

The use of the MMI format has yielded promising results in student admissions to health professions. Compared to traditional interviews, MMIs offer improved reliability and validity, which should be expected of a selection instrument in a high-stakes selection

setting (Patterson et al. 2016). For example, the multiple independent sampling methodology is thought to contribute to its reliability, as one assessment does not influence another (Hanson et al. 2012). Generally, the reliability of MMI has been reported to be acceptable; however, the interpretation of reliability in studies has not always been clearly explained (Pau et al. 2013). Factor analysis is recommended for identifying which stations assess the same attributes (Lemay et al. 2007), and generalisability analysis is often carried out to compute the overall reliability of the MMI considering all sources of variance (Pau et al. 2013), such as interviewer, station, or applicant. In a recent study piloting the MMI procedure in Finnish teacher education, approximately 63% of the variance between the MMI scores could be attributed to the applicant, while measurement error accounted for about 21% of the variance (Metsäpelto, Utriainen, et al. 2022). In most of the stations, interviewer and circuit effects were small (under 10%), while the station explained around 20% of the variance, pointing to varying levels of difficulty between stations.

There has been disagreement about whether MMIs can measure only one attribute, one attribute per station, or one attribute per social skill dimension, regardless of station (Breil et al. 2020). For example, measures of different noncognitive attributes assessed within the same station have been found to be highly correlated, which might limit the interpretability of MMI scores (Oliver et al. 2014). This has also led to the common practice of reporting total scores within each station instead of separate attribute-based scores. Thus, it remains unclear whether MMI scores are able to capture one or multiple attributes. Oliver et al. (2014) reported evidence supporting the measurement of multiple attributes within one station but highlighted the need to clearly define the aspects of the noncognitive attributes that are intended to be assessed. Additionally, studies by Lemay et al. (2007) and Hecker et al. (2009) provided similar evidence supporting the multidimensionality of MMI, that is, the ability of the format to assess multiple noncognitive attributes.

In medical education, consistent evidence has been reported concerning the predictive validity of MMIs. According to a systematic review (Patterson et al. 2016), performance on MMIs was related to subsequent performance on both undergraduate and postgraduate examinations of clinical competence, as well as on other examinations. Kelly et al. (2014) proved the opposite, showing that MMI was not highly predictive of first-year success in the programme. However, they suggested that predictive validity patterns may change as students progress in their studies and gain more clinical experience. In the teacher education context, Salingré and MacMath (2021) reported few correlations between MMI and success measures in the long practicum, suggesting the need for more research.

Concerning the discriminant validity of MMI, the format has shown low or non-existent associations with traditional admission tools, such as cognitive ability, pre-entry academic qualifications (e.g. GPA), or interview (e.g. Eva et al. 2004; Hecker et al. 2009; Pau et al. 2013; Rees et al. 2016). However, scores from MMIs with only interview stations were strongly associated with scores from MMIs that included different kinds of exercises, demonstrating convergence between different task formats (Gafni et al. 2012). According to Breil et al. (2020), most previous studies have focused on the relationships between overall MMI scores and other selection measures, while it would be important to consider the correlation between individual stations and other selection measures if MMI is assumed to measure several distinct attributes.

In the medical context, several studies indicate good face validity for MMI, characterising it as a more acceptable method of admission assessment than traditional interviews by both interviewers and interviewees (e.g. Dore et al. 2010; Kelly et al. 2014). In teacher education, MMI has also been perceived positively by both applicants and interviewers (Metsäpelto, Utriainen, et al. 2022). The special advantage of MMIs is that applicants can improve their performance at the next station, regardless of possible failures in previous stations (Pau et al. 2013; Rosenfeld et al. 2008). Several studies on the acceptability of MMIs state that MMI scores are not related to applicants' gender, disadvantages, previous school success, or access to paid coaching (Pau et al. 2013; Uijtdehaage, Doyle, and Parker 2011). However, some studies have reported that applicants of aboriginal or rural background receive lower scores compared to other applicants (Rees et al. 2016). Acceptability can be increased by combining different types of questions within the stations (Yamada et al. 2017).

The current study: MMI as a part of student admission to teacher education in Finland

In Finland, students are selected for ITE via a two-phase selection process. The first phase includes the preselection of applicants based on their weighted matriculation exam scores, which are applied to a percentage (e.g. 60%) of applicants. They are granted the right to enter the second phase without participating in the first phase screening. Applicants without matriculation exams or high enough scores may take first phase screening, which aims to measure the applicants' cognitive competencies, such as conceptual comprehension. It is comprised of a nationally planned and coordinated written exam (VAKAVA) that comprises reading scientific texts and answering multiple-choice questions based on the texts within a time limit. Based on the success in the first phase, at least twice as many applicants compared to the number of study places are invited in the second phase.

The final screening of applicants takes place in the second phase, which focuses on applicants' noncognitive attributes, such as social skills. The final screening is based on a nationally planned and coordinated MMI format in which applicants move through a sequence of stations with different interviewers and respond to structured tasks (Eva et al. 2004). The attributes measured in the MMI are derived from the multidimensional adapted process model (MAP) of teaching (Metsäpelto, Poikkeus, et al. 2022), highlighting applicants' suitability for the teaching profession. In the final selection decision, neither the scores from the preselection nor the first phase screening are considered; only the scores of the MMI are used to rank the students.

In both the first and second phase screenings, the procedures and scoring are uniform across universities, and both phases apply to all teacher education programmes in Finland. The applicants can simultaneously apply to several programmes within the same or multiple universities, but they must rank their choices and can only enter one programme in which they score higher than its cut-off.

The current study addressed questions concerning the reliability as well as construct and discriminant validity of MMI as a part of student admission to teacher education programmes in Finland. The MMI format was piloted in some teacher education units in

2019, after which it was further developed; since 2020, it has been utilised nationwide. Since the format has been fairly recently applied to the teacher education context but is less studied within it, the study aimed to shed more light on the matter by answering the following research questions:

In the context of student admission to teacher education,

(1) Does the MMI format demonstrate reliability?
   - How much of the variation in MMI scores in each station is explained by the clustering of applicants to different interviewers and circuits?
   - How much of the variation in MMI scores is attributable to the interviewer, circuit, station, applicant, and measurement error?
(2) Does the MMI format demonstrate validity?
   - Does the MMI measure several noncognitive attributes, as it is designed to do?
   - To what extent does MMI assess attributes that are not assessed by other admission variables (matriculation exam scores and VAKAVA scores)?

The first research question addressed the reliability of the MMI in terms of how much of the variance between the scores could be attributed to the participants and how much to the other factors. In a previous study, the largest variance component in the MMI total score was for the applicant, but it reflected only marginally satisfactory reliability (Metsäpelto, Utriainen, et al. 2022). At the time of the earlier study, the MMI procedure was piloted for the first time in Finnish teacher education admissions, whereas at the time of the current study, the format had become established. Thus, we expected slightly higher reliability in terms of applicant scores.

The second research question concerning the validity of the MMI was divided into two sub-questions. The first aimed to examine whether the data structure of the MMI scores supports the presence of multiple noncognitive attributes as independent test dimensions. Given the explicit measurement and supposed distinctiveness of five MMI dimensions, we expected a stronger model fit for a five-factor solution than for a one-factor solution, suggesting that only one attribute is measured; or for a four-factor solution, suggesting that each station would measure only one target attribute (instead of the two attributes that were aimed at; see Oliver et al. 2014).

The second sub-question pertained to the relationship between the MMI and the other admission measures. Based on earlier research, we assumed that the MMI measures something other than the cognitive skills measured by the other admission variables, namely, matriculation exam scores and VAKAVA scores (for GPA, see Eva et al. 2004, 2012; Reiter et al. 2007). Thus, we expected only modest correlations between MMI station scores and other admission measures.

Materials and methods

Participants

The participants of the study comprised the entire population of applicants ($N = 3309$) who had applied to at least one of the three largest teacher education programmes in Finland (primary teacher education, early childhood teacher education, or special

needs teacher education) and had participated in the second phase of student admission (the aptitude test organised in the MMI format in 2021). The data were drawn from the national register for study programmes leading to a degree, maintained by the Finnish National Agency for Education. Complying with the European General Data Protection Regulation, the participants were informed about the processing of their personal data and given the opportunity to refuse participation. Three applicants requested non-participation in the study, resulting in a final sample size of 3306.

Measures and procedure

The multiple mini-interview

The multiple mini-interview (MMI) procedure was designed based on a pilot study (Metsäpelto, Utriainen, et al. 2022), where five attributes were measured through five five-minute stations. However, due to resource-intensiveness and a large number of applicants compared to many studies in the medical field (e.g. Eva et al. 2004; Uijtdehaage, Doyle, and Parker 2011), in the current study, the MMI circuit was condensed into four stations, each lasting five minutes, with a three-minute turnaround between stations. While the number of stations was smaller than reported in several studies (see, e.g. the reviews of Knorr and Hissbach 2014; Pau et al. 2013), it is not exceptionally small, as even three-station MMIs have been used (e.g. Klassen and Kim 2021; Onyon, Wall, and Goodyear 2009). Each applicant rotated through the four-station circuit and met with a different interviewer at each station. The MMIs were organised in traditional face-to-face mode (Kok et al. 2021) and synchronously at different universities in Finland, totalling 51 circuits and 204 interviewers. All the interviewers received one-day training on the aims and implementation of MMIs as well as on the administration and scoring of their stations.

The stations were designed by an admission committee of eight senior staff members from different universities. Station development was based on the MAP framework (Metsäpelto, Poikkeus, et al. 2022), which specifies the key competence domains perceived to be critical for the teaching profession. The model considers high-quality teaching learner-centred and constructivist, and emphasises teaching interactions, students' active role in learning, and teachers' emotional and learning support for students. Stations were designed according to certain attributes that were considered crucial in developing these skills and that indicated applicants' general suitability for the teaching profession.

Each of the four stations focused on one target attribute (motivation for pursuit of a teaching career, skills in managing emotions, intercultural competence, and social problem solving), which was assessed with a criterion-referenced scoring rubric. Additionally, social skills were assessed at each station. The stations included subtasks that comprised different formats and content types, such as reflection on certain topics. At each station, applicants could earn a maximum of 24 points on the measured attribute, and the same total score applied to the attribute of social skills assessed at each station (with a maximum of six points at each). Thus, the maximum MMI score was 120. The total score of the applicant and attribute-related scores were utilised in the analyses. For the first station, however, a total score of 18 was applied, since the first subtask of the station was not related to the main target attribute of the station (i.e. motivation), but to

educability, and was thus theoretically separate from the other subtasks. If the applicant failed on the first subscale of the task, the entire MMI failed.

The matriculation exam

The matriculation exam is the final exam of upper secondary school in Finland. It is completed when a student has passed at least four exams at the baccalaureate level at the end of the studies, which typically last three years. The separate exam grades of only the mother tongue and mathematics (basic and advanced syllabi) were used in this study. The Latin exam grades were converted to numeric values from 0 (improbatur; failed test) to 7 (laudatur; outstanding). In addition to separate grades of mother tongue and mathematics, the total of the weighted matriculation exam scores was used in the analyses, as it is utilised in the preselection, by which 60% of applicants are chosen to go straight to MMI without having to participate in VAKAVA. The total score was calculated based on four subjects: mother tongue, mathematics (basic or advanced studies), one subject of humanities and sciences, and one language (short, average, or extended studies). The maximum weighted matriculation exam score was 123.9.

The VAKAVA written entrance exam

The VAKAVA written entrance exam is a multiple-choice exam based on research papers and related multiple-choice questions (see Haataja et al. 2023). Using the available articles, the participants had to answer 12 tasks, totalling 106 multiple-choice items. They were awarded one point for correct answers and lost points for incorrect answers. The maximum score was 106.

Analysis

First, descriptive statistics and correlations of the MMI attribute scores (i.e. station scores and social skills scores measured separately for each station) and the MMI total score (i.e. the sum of attribute scores) were calculated to provide an overview of the applicant scores. Since the attribute scores were not normally distributed, nonparametric Spearman correlations were used.

To calculate the reliability of the MMI, a similar approach to Metsäpelto, Utriainen, et al. (2022) was utilised using Mplus 8.4 (Muthén and Muthén 1998–2017). First, intraclass correlations (ICC) were calculated to examine how much of the variation in applicant scores was explained by the clustering of applicants to different interviewers (i.e. interviewer effect) and to the four-station circuits (i.e. circuit effect). ICC is considered a desirable measure of reliability, as it reflects both the degree of correlation and agreement between measurements (Koo and Li 2016). Furthermore, it measures the relatedness of observations within a cluster, ranging from 0 to 1, where 0 indicates a non-existent correlation between observations within a cluster, and 1 indicates identity among all observations within a cluster (Killip, Mahfoud, and Pearce 2004).

Second, we analysed how much of the variation in total scores in MMI was attributable to the interviewer, circuit, station, applicant, and measurement error. Following the idea of generalisability theory (e.g. Brennan 2010), the contributions of different factors (i.e. variance components) to the variance of the MMI total score were estimated. A multilevel

model was utilised with three factors to estimate the variance component for the interviewer and the circuit for which the applicant was assigned, as well as the station, to examine the mean differences in scores obtained by applicants at each station (e.g. Marsh, Martin, and Cheng 2008).

In the current study, the outcome was the attribute scores. The first station scores (0–18) were scaled similarly to match the other scores (0–24) before modelling. The stations were treated as fixed effects in the model as the applicant moved through four stations measuring different variables. Dummy variables were used to calculate the stations' relative difficulty level effects, which were regressed on the applicants' scores at each station. However, each applicant attended one of the 51 circuits and was assessed by the same interviewer at each station with regard to the applicant's success at the station's target attribute, as well as the social skills measured at each station. The remaining variance in the model was residual variance, which could not be attributed to any of the variables in the model. Residual variance had two sources of variation, the applicant's true score variation and measurement error, which the model did not separate out. Thus, the variance attributable to the applicants was calculated otherwise.

As each station and social skills measurement comprised two to six subscales, Cronbach's alphas were calculated for each station attribute and social skills measured at each station. As a measure of internal consistency, Cronbach's alpha provided an estimate of how much of the variation in the attribute scores was explained by the applicant's true scores and how much was attributable to measurement error. The sum of the attribute-specific measurement errors was used as an overall indicator of the amount of measurement error in the total MMI score and was subtracted from the residual, resulting in the estimate of the true score variance for the applicants (see Metsäpelto, Utriainen, et al. 2022).

Third, confirmatory factor analyses (CFA) were applied to test the underlying factor structure of the MMI. Since two attributes were measured by the same interviewer at each station, residual covariances were allowed between the station's target attribute and the social skills scores within each station. The CFA models were estimated using full information maximum likelihood estimation with robust standard errors (MLR), which can handle non-normal data. Model fit was evaluated using the chi square test and fit indices with general cut-offs (see Schreiber et al. 2006): the chi square test (nonsignificant p value indicated a good fit), the comparative fit index (CFI), and the Tucker – Lewis index (TLI; values .95 or above indicate acceptable fit), the root mean square error of approximation (RMSEA; values below .05 indicate a good fit), and the standardised root mean square residual (SRMR; values below .08 indicate a good fit). Additionally, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) were used for the model comparison (the smaller the better). Finally, the MMI scores were examined in relation to other admission variables (matriculation exam scores and VAKAVA scores) using Spearman correlations.

Table 1. MMI attribute and total score descriptive statistics and Spearman correlations between the scores.

	All applicants ($N = 3306$)				Correlations (r_s)					
Score/attribute	M	SD	SK	Rku	1.	2.	3.	4.	5.	6.
1. Station 1: Motivation for pursuing a teaching career	18.57	4.06	−.73	.61	—					
2. Station 2: Skills in emotion management	18.85	3.13	−.40	.59	.20***	—				
3. Station 3: Intercultural competence	13.52	3.77	.02	.00	.29***	.20***	—			
4. Station 4: Social problem-solving	18.72	2.12	−1.56	9.65	.22***	.16***	.23***	—		
5. All stations: Social skills	22.08	2.12	−2.30	10.69	.45***	.28***	.32***	.32***	—	
6. MMI total score	91.74	10.21	−1.06	4.71	.72***	.57***	.58***	.50***	.65***	—

Note. The scores for each station ranged from 0 to 24. ***$p < .001$.

Results

Descriptive statistics and correlations of MMI attribute scores

As shown in Table 1, there was a relatively large variation in the MMI attribute scores. On average, the applicants scored the highest on social skills measured at each station, and the lowest on intercultural competence. Correlations between the attributes ranged from low to moderate (.16–.45), suggesting that different attributes were measured at different stations. Correlations were the highest between social skills and motivation to pursue a teaching career (.45). Overall, the social skills scores measured at each station correlated the highest with the other attribute scores, which might be explained by the fact that they were rated by the same interviewer as each station score.

Interviewer and circuit effects

ICCs were utilised to determine how much variation in applicant scores at each station could be explained by the clustering effect of the interviewer and the circuit (Table 2). ICCs at all stations remained fairly low (0.08–0.17), showing that nearly all the measured variance in these attributes was related to sources of variance other than factors concerning the interviewer or circuit. When the circuit was used as a clustering variable, the results remained similar: the four-station circuit to which the applicant was assigned had a very small effect on the variance in the MMI total scores.

Table 2. Intraclass correlations (ICC) of the five MMI target attributes and total scores.

		95% Confidence interval	
Clustering variable	ICC	Lower bound	Upper bound
Interviewer			
Station 1	.11***	.07	.15
Station 2	.08***	.04	.11
Station 3	.17***	.09	.25
Station 4	.10***	.06	.14
Social Skills	.12***	.12	.17
4-station circuit			
Social Skills (all stations)	.09***	.06	.12
Total MMI Score	.08***	.05	.12

***$p < .001$.

Table 3. Variance components for MMI total scores.

	MMI total score		Applicant to applicant variations in total score	
Source of variance	Variance component	% of total variance	Variance component	% of total variance
Interviewer	1.44	7.5	1.44	10.1
Circuit	0.22	1.1	0.22	1.5
Station	4.89	25.4	-	-
Residual	12.68	65.9	12.68	88.4
Total	19.23	100.0	14.34	100.0

Table 4. Cronbach's alpha reliabilities, variances, and variances of measurement error.

Station/Attribute		No. of subscales at a station	N	Variance	Cronbach's alpha	Variance of measurement error
1	Motivation for pursuing a teaching career	4	3306	16.49	0.77	3.79
	Social skills	2	3306	0.72	0.71	0.21
2	Skills in emotion management	6	3306	9.78	0.43	5.55
	Social skills	2	3306	0.60	0.60	0.24
3	Intercultural competence	3	3306	14.25	0.77	3.32
	Social skills	2	3306	0.63	0.69	0.20
4	Social problem-solving	4	3306	4.48	0.54	2.08
	Social skills	2	3306	0.75	0.57	0.32
Total		25				15.72

Variance components

Table 3 shows the results of the trilevel modelling, illustrating the variance components of the MMI total score. First, we estimated all sources of variance (interviewer, circuit, and station) that explained the applicant's MMI total score. The variance in the MMI total score explained by the circuit was minimal (1.1%) and by the interviewer fairly small (7.5%), whereas the station explained approximately one-fourth of the variance (25.4%). Thus, there were relatively small differences in the MMI total scores as a function of the specific circuit or the interviewer to which the applicant was assigned, whereas each station explained a significant portion of the variation. This could be interpreted as highlighting the differences between stations: one applicant could get very different scores from different stations, pointing to the fact that the stations measured different attributes. The remaining residual variance (65.9%) was considered to represent the variance concerning the applicant's true score and measurement error. All variance components were statistically significant.

Table 3 also illustrates the estimated sources of variance, explaining the applicant-to-applicant variations in the MMI total score. Since each applicant moved through the same stations, the relative difficulty of the stations did not generate variation between applicants and was removed from the sources of variation. Thus, the applicant-to-applicant variation in the MMI total score was explained by the interviewer (10.1%), the circuit (1.5%), and the residual (88.4%).

To differentiate between the true score variance related to the applicant and the measurement error variance from the residual variance, the total measurement error was calculated for each attribute score using Cronbach's alphas (Table 4), resulting in a total of 15.72. As the MMI total score variance was 104.19, the proportion of the total measurement error from the total score variance was 15.09%. Then, the total

Table 5. Model fit of the CFA models.

Model	χ2 (df), *p*	CFI, TLI	RMSEA	SRMR	AIC, BIC
1-factor model	4231.43 (177), .000	.74, .66	.08	.08	157995.25, 158593.39
4-factor model	738.73 (171), .000	.96, .95	.03	.03	154166.79, 154801.55
5-factor model	**572.00 (167), .000**	**.97, .96**	**.03**	**.02**	**153994.72, 154653.90**

Note. χ2 (df) = chi-square test of model fit with degrees of freedom and p-value, CFI = comparative fit index, TLI = Tucker – Lewis fit index, RMSEA = root mean square error of approximation, SRMR = standardised root mean square residual, AIC, BIC = information criteria.

measurement error variance was subtracted from the variance component of the residual variance (88.4%; see Table 4), resulting in 73.3% for the applicants' true score variance.

Dimensionality of MMI scores

Next, the proposed underlying factor structure was analysed using CFAs. To test the structure of the MMI, three different CFA models were applied. The first model was a one-factor model, based on the hypothesis that MMI could not measure various noncognitive attributes but that all the tasks would measure a general MMI score or attribute. The second model included four factors, suggesting that the tasks at each station would measure one target attribute. The third model comprised a five-factor solution that matched the theoretical basis of the applied MMI format, suggesting that five target attributes could be measured with four stations. In all models, three subtasks (1a – c) at Station 2 had to be omitted from the solutions due to low factor loadings. These three subtasks applied a different format from the other subtasks of the station, which might explain their poor loadings. Additionally, residual covariances were permitted between the social skills scores and other attribution scores measured at the same station (Oliver et al. 2014). The covariances seemed justified since the scorings were made by the same interviewer. As can be seen from Table 5, CFA with five factors resulted in the best fit. Furthermore, the item loadings were good (.37–.82), and there were no cross-loadings (see Table 6). Thus, CFA supported the intended structure of the MMI. Cronbach's alpha and McDonald's omega values showed acceptable internal consistency for motivation to pursue a teaching career, intercultural competence, and social skills, whereas the internal consistency of skills in emotion management and social problem-solving was poor.

Relationships between MMI scores and other admission variables

The correlations between the other admission variables and the MMI dimensions were low (Table 7). The highest correlations emerged between the applicants' VAKAVA scores and the MMI attributes (.10–21), as well as the MMI total score (.25). Thus, it seems that the MMI and other admission variables measured different attributes.

Discussion

The aim of the current study was to shed more light on the reliability and validity of the MMI format as a part of admission to teacher education using an extensive sample, the entire cohort of applicants. While the format is widely used and studied in student admission to health professions, there is a paucity of research concerning the format in

Table 6. Factor structure and loadings of MMI scores.

Station	Factor	M	SD	Factor loading	S. E.	Cronbach's Alpha	McDonald's Omega
1	Motivation for pursuing a teaching career					.77	.81
	• Subtask 2	5.21	1.36	.69	.01		
	• Subtask 3a	3.54	1.15	.83	.01		
	• Subtask 3b	2.46	.64	.82	.01		
	• Subtask 3c	2.73	.57	.58	.02		
2	Skills in emotion management					.52	.53
	• Subtask 2	5.20	1.62	.48	.02		
	• Subtask 3a	4.18	1.37	.62	.02		
	• Subtask 3b	4.05	1.12	.48	.02		
3	Intercultural competence					.77	.77
	• Subtask 1	4.78	1.36	.72	.02		
	• Subtask 2	4.53	1.51	.70	.01		
	• Subtask 3	4.21	1.69	.76	.01		
4	Social problem-solving					.54	.54
	• Subtask 1	4.70	.59	.40	.04		
	• Subtask 2	4.24	.92	.49	.02		
	• Subtask 3	4.67	.83	.53	.03		
	• Subtask 4	5.11	.90	.49	.03		
All	Social skills					.72	.70
1	• Subtask 1	2.83	.44	.38	.03		
	• Subtask 2	2.72	.52	.37	.03		
2	• Subtask 1	2.81	.42	.46	.03		
	• Subtask 2	2.73	.49	.49	.03		
3	• Subtask 1	2.78	.46	.46	.03		
	• Subtask 2	2.80	.44	.49	.03		
4	• Subtask 1	2.75	.48	.46	.03		
	• Subtask 2	2.65	.55	.52	.03		

Table 7. Spearman's correlations between MMI attributes and other admission variables.

	N	*M*	*SD*	1	2	3	4	5	MMI total score
Weighted matriculation exam scores	3279	56.87	24.22	−.01	.07***	.02	.05**	.01	.05*
Math, advanced	895	4.09	1.24	−.03	−.03	−.07*	.08*	−.03	−.02
Math, basic	1605	4.20	1.39	.04	.06*	.06*	.06*	.04	.08**
Mother tongue	3015	4.53	1.19	.12***	.08***	.10***	.06**	.12***	.15***
VAKAVA scores	2360	47.82	17.04	.19***	.10***	.19***	.17***	.21***	.25***

Note. 1 Motivation for pursuing a teaching career, 2 Skills in emotion management, 3 Intercultural competence, 4 Social problem-solving, 5 Social skills. *$p < .05$.**$p < .01$.***$p < .001$.

teacher education admission, where it has only recently been applied (Metsäpelto, Utriainen, et al. 2022; Salingré and MacMath 2021).

In admission to Finnish teacher education programmes, the second phase of applicant screening is currently based on the MMI format. In the year of the study, the format included five noncognitive target attributes measured at four stations. Each station included a separate target attribute and an attribute that was measured at each station. The attributes were derived from the MAP model of teacher competence (Metsäpelto, Poikkeus, et al. 2022) based on an extensive literature review on teaching.

The analyses concerning the reliability of the MMI showed that the clustering of applicants to different interviewers explained a fairly small amount of variance for most of the measured attributes (around 10%). This shows that the applicants were treated

reliably and consistently in the assessment of most attributes. Thus, the scores given by one interviewer did not resemble each other more than the scores of other applicants by other interviewers (Metsäpelto, Utriainen, et al. 2022). At Station 3, targeting applicants' intercultural competence, the intraclass correlation was somewhat higher (17%), hinting that interviewer bias may have affected the scoring to some degree.

The reliability of the MMI was further studied using multilevel modelling. When the sources for variance in the MMI total score were studied, the effect of the circuit was minimal (1.1%) and that of the interviewer was fairly small (<8%). However, the station explained about one-fourth of the overall variance, highlighting that one applicant received different scores from different stations. This seems understandable, as the stations were designed to measure different noncognitive attributes; thus, their perceived difficulty for the applicant probably varied. Therefore, one could score high on social problem-solving but low on intercultural competence, for instance. These results are very much in line with the previous study in the context of Finnish teacher education (Metsäpelto, Utriainen, et al. 2022).

The applicants' true score variance turned out to be 73.3%, which can be considered acceptable (e.g. Dore et al. 2010) and shows improvement from the earlier study when the MMI procedure was piloted (63.3%; Metsäpelto, Utriainen, et al. 2022). Although the MMIs are redesigned every year and thus are not directly comparable, this can be seen as a positive development, indicating that the reliability of the procedure has been strengthened by experience. With only four 5-minute stations and five measured attributes aiming at cost-effective implementation, the results concerning reliability are encouraging. However, the stations' internal consistency varied remarkably (Cronbach's alphas 0.53–0.77), and despite the structured stations and criterion-based scoring rubric, approximately 15% of the variation in the MMI total score was still attributable to measurement error. Intra-station reliability is usually expected to be high (Pau et al. 2013), and some studies report results compliant with this assumption (e.g. Callwood et al. 2018; Dore et al. 2010; Lemay et al. 2007). However, intra-station (or domain) reliability is not often studied, and other examples of variability exist, such as in Dowell et al. (2012). The reasons for inconsistency might be related to different task types applied in the stations, as well as different number of subtasks within stations. These findings emphasise that more attention should be paid to developing the stations and their subtasks.

Our results support the assumption that the MMI is able to assess and differentiate among multiple noncognitive attributes (Hecker et al. 2009; Lemay et al. 2007; Oliver et al. 2014). The correlation analyses showed a relatively low correlation between the target dimensions, suggesting that the stations measure different attributes. However, the correlations between social skills dimensions, measured at every station, and each station attribute were the highest, suggesting that the dimension of social skills was the least distinguishable attribute of the five measured attributes. However, CFAs supported the five-factor structure over the one-factor model, suggesting that MMI would measure only one target attribute as well as over the four-factor model, where the dimension of social skills was embedded in the other four target attributes and not as a separate dimension.

When the MMI scores were contrasted with the cognitive measures of matriculation examination scores and VAKAVA written entrance examination scores, only weak to nonexistent associations between them were found. This provides evidence that the MMI measures something other than the cognitive measures of the application phase (e.g. Eva

et al. 2004, 2012; Reiter et al. 2007). Thus, the results support the role of MMI in the admission process as a supplementary measure for cognitive screening.

The strength of the study lies in the sophisticated statistical modelling based on representative data from the entire yearly population of applicants to the largest teacher education programmes in Finland. However, one needs to consider the one-off nature of the MMI: since the stations are redesigned every year, their comparability to earlier and future studies is limited. However, the measured attributes show a certain consistency from year to year (cf. Metsäpelto, Utriainen, et al. 2022), as they are based on the MAP model (Metsäpelto, Poikkeus, et al. 2022). Further, although the study explored the reliability and validity of the MMI in versatile ways, it was in no account comprehensive. Future studies should therefore be conducted, for example, to assess its predictive validity, that is, whether the MMI scores are associated with study achievement in teacher education as well as success and persistence in the teacher profession, or convergence validity, that is, whether other tests measuring similar attributes give similar results. Additionally, the study did not consider the fairness of the MMI to all applicants, that is, whether applicants' age, gender, socioeconomic status or ethnic background had an effect on their performance (cf. Rees et al. 2016).

Conclusion

In conclusion, our study provides encouraging support for the reliability and validity of using the MMI format as a part of student admission to teacher education. However, more research and development work are needed to improve the internal consistency of the stations. Considering the resource intensity of the traditional face-to-face MMI, other solutions could also be considered, such as an automated interview grounded in MMI methodology piloted in healthcare student selections (Callwood et al. 2022).

Acknowledgments

The authors wish to thank the Research Council of Finland [#342191] for the financial support for conducting this study. Furthermore, the authors wish to thank all the participants in the study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Research Council of Finland under Grant [#342191]. However, the funding agency had no involvement in study design, in the collection, analysis, and interpretation of data, nor writing the report or the decision to submit the article for publication.

Notes on contributors

Henna Vilppu PhD (education), works as a University Research Fellow at the Department of Teacher Education, University of Turku, Finland. Her research interests concern teaching and learning at

different higher education contexts, as well as teacher education and teachers' professional development.

Eero Laakkonen Lic. Econ. (Statistics), works as a senior researcher at the Department of Teacher Education, University of Turku, Finland. His research interests concern the application of quantitative research methods in behavioural sciences, e.g., statistical multivariate methods, structural equations models, statistical computing.

Eeva Haataja PhD (education), works as a university lecturer at the Faculty of Educational Sciences, University of Helsinki, Finland. Her research interests concern teachers' professional vision and teacher-student interaction in mathematics education, as well as teacher education and teachers' professional development.

Asko Tolvanen PhD (Statistics), works as professor in Methodology Centre for Human Sciences, at the University of Jyväskylä, Finland. His research interest concern applying structural equation modelling in the context of multilevel and mixture modelling.

Riitta-Leena Metsäpelto PhD (Psychology) and Adjunct Professor, works as a senior researcher at the Department of Teacher Education, University of Jyväskylä, Finland. Her research interests concern student selection for teacher education, teacher competences and teachers' professional development.

References

Bardach, L., and R. Klassen. 2020. "Smart Teachers, Successful Students? A Systematic Review of the Literature on Teachers' Cognitive Abilities and Teacher Effectiveness." *Educational Research Review* 30:100312. https://doi.org/10.1016/j.edurev.2020.100312.

Bardach, L., R. Klassen, and N. Perry. 2022. "Teachers' Psychological Characteristics: Do They Matter for Teacher Effectiveness, Teachers' Well-Being, Retention, and Interpersonal Relations? An Integrative Review." *Educational Psychology Review* 34 (1): 259–300. https://doi.org/10.1007/s10648-021-09614-9.

Breil, S., B. Forthmann, A. Hertel-Waszak, H. Ahrens, B. Brouwer, E. Schönefeld, B. Marschall, and M. Back. 2020. "Construct Validity of Multiple Mini Interviews—Investigating the Role of Stations, Skills, and Raters Using Bayesian G-Theory." *Medical Teacher* 42 (2): 164–171. https://doi.org/10.1080/0142159X.2019.1670337.

Brennan, R. 2010. "Generalizability Theory and Classical Test Theory." *Applied Measurement in Education* 24 (1): 1–21. https://doi.org/10.1080/08957347.2011.532417.

Callwood, A., D. Cooke, S. Bolger, A. Lemanska, and H. Allan. 2018. "The Reliability and Validity of Multiple Mini Interviews (MMIs) in Values Based Recruitment to Nursing, Midwifery and Paramedic Practice Programmes: Findings from an Evaluation Study." *International Journal of Nursing Studies* 77:138–144. https://doi.org/10.1016/j.ijnurstu.2017.10.003.

Callwood, A., L. Gillam, A. Christidis, J. Doulton, J. Harris, M. Piano, A. Kubacki, et al. 2022. "Feasibility of an Automated Interview Grounded in Multiple Mini Interview (MMI) Methodology for Selection into the Health Professions: An International Multimethod Evaluation." *BMJ Open* 12 (2): e050394. https://doi.org/10.1136/bmjopen-2021-050394.

Cameron, A., L. MacKeigan, N. Mitsakakis, and J. Pugsley. 2017. "Multiple Mini-Interview Predictive Validity for Performance on a Pharmacy Licensing Examination." *Medical Education* 51 (4): 379–389. https://doi.org/10.1111/medu.13222.

D'Agostino, J., and S. Powers. 2009. "Predicting Teacher Performance with Test Scores and Grade Point Average: A Meta-Analysis." *American Educational Research Journal* 46 (1): 146–182. https://doi.org/10.3102/0002831208323280.

Dodson, M., B. Crotty, D. Prideaux, R. Carne, A. Ward, and E. de Leeuw. 2009. "The Multiple Mini-Interview: How Long Is Long Enough?" *Medical Education* 43 (2): 168–174. https://doi.org/10.1111/j.1365-2923.2008.03260.x.

Dore, K., S. Kreuger, M. Ladhani, D. Rolfson, D. Kurtz, K. Kulasegaram, A. Cullimore, et al. 2010. "The Reliability and Acceptability of the Multiple Mini-Interview As a Selection Instrument for Postgraduate Admissions." *Academic Medicine* 85 (10): 60–63. https://doi.org/10.1097/ACM.0b013e3181ed442b.

Dowell, J., B. Lynch, H. Till, B. Kumwenda, and A. Husbands. 2012. "The Multiple Mini-Interview in the UK Context: 3 Years of Experience at Dundee." *Medical Teacher* 34 (4): 297–304. https://doi.org/10.3109/0142159X.2012.652706.

Eva, K., H. Reiter, J. Rosenfeld, K. Trinh, T. Wood, and G. Norman. 2012. "Association Between a Medical School Admission Process Using the Multiple Mini-Interview and National Licensing Examination Scores." *Journal of the American Medical Association* 308 (21): 2233–2240. https://doi.org/10.1001/jama.2012.36914.

Eva, K., J. Rosenfeld, H. Reiter, and G. Norman. 2004. "An Admissions OSCE: The Multiple Mini Interview." *Medical Education* 38 (3): 314–326. https://doi.org/10.1046/j.1365-2923.2004.01776.x.

Gafni, N., A. Moshinsky, O. Eisenberg, D. Zeigler, and A. Ziv. 2012. "Reliability Estimates: Behavioural Stations and Questionnaires in Medical School Admissions." *Medical Education* 46 (3): 277–288. https://doi.org/10.1111/j.1365-2923.2011.04155.x.

Haataja, E., A. Tolvanen, H. Vilppu, M. Kallio, J. Peltonen, and R.-L. Metsäpelto. 2023. "Measuring Higher-Order Cognitive Skills with Multiple-Choice Questions: Potential and Pitfalls of Finnish Teacher Education Entrance." *Teaching and Teacher Education* 122:103943. https://doi.org/10.1016/j.tate.2022.103943.

Hanson, M., K. Kulasegaram, D. Coombs, and J. Herold. 2012. "Admissions File Review: Applying the Multiple Independent Sampling (MIS) Methodology." *Academic Medicine* 87 (10): 1335–1340. https://doi.org/10.1097/ACM.0b013e3182674629.

Hecker, K., T. Donnon, C. Fuentealba, D. Hall, O. Illanes, D. Morck, and C. Muelling. 2009. "Assessment of Applicants to the Veterinary Curriculum Using a Multiple Mini-Interview Method." *Journal of Veterinary Medical Education* 36 (2): 166–173. https://doi.org/10.3138/jvme.36.2.166.

Kelly, M., J. Dowell, A. Husbands, J. Newell, S. O'Flynn, T. Kropmans, F. Dunne, and A. Murphy. 2014. "The Fairness, Predictive Validity and Acceptability of Multiple Mini Interview in an Internationally Diverse Student Population—A Mixed Methods Study." *BMC Medical Education* 14 (1): 267. https://doi.org/10.1186/s12909-014-0267-0.

Killip, S., Z. Mahfoud, and K. Pearce. 2004. "What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers." *The Annals of Family Medicine* 2 (3): 204–208. https://doi.org/10.1370/afm.141.

Klassen, R., and L. Kim. 2021. *Teacher Selection: Evidence-Based Practices*. Wiesbaden: Springer.

Klassen, R., L. Kim, J. Rushby, and L. Bardach. 2020. "Can We Improve How We Screen Applicants for Initial Teacher Education?" *Teaching and Teacher Education* 87:102949. https://doi.org/10.1016/j.tate.2019.102949.

Knorr, M., and J. Hissbach. 2014. "Multiple Mini-Interviews: Same Concept, Different Approaches." *Medical Education* 48 (12): 1157–1175. https://doi.org/10.1111/medu.12535.

Kok, K., L. Chen, F. Irdayati Idris, N. H. Mumin, H. Ghani, I. Nazurah Zulkipli, and M. Ann Lim. 2021. "Conducting Multiple Mini-Interviews in the Midst of COVID-19 Pandemic." *Medical Education Online* 26 (1): 1. https://doi.org/10.1080/10872981.2021.1891610.

Koo, T., and M. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Lehane, P., Z. Lysaght, and M. O'Leary. 2023. "A Validity Perspective on Interviews as a Selection Mechanism for Entry to Initial Teacher Education Programmes." *European Journal of Teacher Education* 46 (2): 293–307. https://doi.org/10.1080/02619768.2021.1920920.

Lemay, J.-F., J. Lockyer, T. Collin, and K. Brownell. 2007. "Assessment of Non-Cognitive Traits Through the Admissions Multiple Mini-Interview." *Medical Education* 41 (6): 573–579. https://doi.org/10.1111/j.1365-2923.2007.02767.x.

Marsh, H., A. Martin, and J. Cheng. 2008. "A Multilevel Perspective on Gender in Classroom Motivation and Climate: Potential Benefits of Male Teachers for Boys?" *Journal of Educational Psychology* 100 (1): 78–95. https://doi.org/10.1037/0022-0663.100.1.78.

Metsäpelto, R.-L., A.-M. Poikkeus, M. Heikkilä, J. Husu, A. Laine, K. Lappalainen, M. Lähteenmäki, et al. 2022. "A Multidimensional Adapted Process Model of Teaching." *Educational Assessment, Evaluation and Accountability* 34 (2): 143–172. https://doi.org/10.1007/s11092-021-09373-9.

Metsäpelto, R.-L., J. Utriainen, A.-M. Poikkeus, J. Muotka, A. Tolvanen, and A. Warinowski. 2022. "Multiple Mini Interview As a Selection Tool for Initial Teacher Education Admissions." *Teaching and Teacher Education* 113:103660. https://doi.org/10.1016/j.tate.2022.103660.

Muthén, L., and B. Muthén. 1998–2017. *Mplus User's Guide*. 8th ed. Los Angeles, CA: Muthén & Muthén.

Oliver, T., K. Hecker, P. Hausdorf, and P. Conlon. 2014. "Validating MMI Scores: Are We Measuring Multiple Attributes?" *Advances in Health Sciences Education* 19 (3): 379–392. https://doi.org/10.1007/s10459-013-9480-6.

Onyon, C., D. Wall, and H. Goodyear. 2009. "Reliability of Multi-Station Interviews in Selection of Junior Doctors for Specialty Training." *Medical Teacher* 31 (7): 665–667. https://doi.org/10.1080/01421590802578236.

Patterson, F., A. Knight, J. Dowell, S. Nicholson, F. Cousans, and J. Cleland. 2016. "How Effective Are Selection Methods in Medical Education? A Systematic Review." *Medical Education* 50 (1): 36–60. https://doi.org/10.1111/medu.12817.

Pau, A., K. Jeevaratnam, Y. Chen, A. Fall, C. Khoo, and V. Nadarajah. 2013. "The Multiple Mini-Interview (MMI) for Student Selection in Health Professions Training: A Systematic Review." *Medical Teacher* 35 (12): 1027–1041. https://doi.org/10.3109/0142159X.2013.829912.

Rees, E., A. Hawarden, G. Dent, R. Hays, J. Bates, and A. Hassell. 2016. "Evidence Regarding the Utility of Multiple Mini-Interview (MMI) for Selection to Undergraduate Health Programs: A BME Systematic Review." *Medical Teacher* 38 (5): 443–445. https://doi.org/10.3109/0142159X.2016.1158799.

Reiter, H., K. Eva, J. Rosenfeld, and G. Norman. 2007. "Multiple Mini-Interviews Predict Clerkship and Licensing Examination Performance." *Medical Education* 41 (4): 378–384. https://doi.org/10.1111/j.1365-2929.2007.02709.x.

Roberts, C., M. Walton, I. Rothnie, J. Crossley, P. Lyon, K. Kumar, and D. Tiller. 2008. "Factors Affecting the Utility of the Multiple Mini-Interview in Selecting Candidates for Graduate-Entry Medical School." *Medical Education* 42 (4): 396–404. https://doi.org/10.1111/j.1365-2923.2008.03018.x.

Rosenfeld, J., H. Reiter, K. Trinh, and K. Eva. 2008. "A Cost Efficiency Comparison Between the Multiple Mini-Interview and Traditional Admissions Interviews." *Advances in Health Sciences Education* 13 (1): 43–58. https://doi.org/10.1007/s10459-006-9029-z.

Salingré, B., and S. MacMath. 2021. "Using Multiple-Mini-Interviews for Admission into Teacher Education." *Canadian Journal of Education* 44 (1): 150–173. https://doi.org/10.53967/cje-rce.v44i1.4401.

Schreiber, J., A. Nora, F. Stage, E. Barlow, and J. King. 2006. "Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review." *The Journal of Educational Research* 99 (6): 323–338. https://doi.org/10.3200/JOER.99.6.323-338.

Uijtdehaage, S., L. Doyle, and N. Parker. 2011. "Enhancing the Reliability of the Multiple Mini-Interview for Selecting Prospective Health Care Leaders." *Academic Medicine* 86 (8): 1032–1039. https://doi.org/10.1097/ACM.0b013e3182223ab7.

Yamada, T., J. Sato, H. Yoshimura, T. Okubo, E. Hiraoka, T. Shiga, T. Kubota, et al. 2017. "Reliability and Acceptability of Six Station Multiple Mini-Interviews: Past-Behavioural versus Situational Questions in Postgraduate Medical Admission." *BMC Medical Education* 17 (1): 57. https://doi.org/10.1186/s12909-017-0898-z.