Teaching and Teacher Education 113 (2022) 103660

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate

Research paper

Multiple mini-interviews as a selection tool for initial teacher education admissions

Riitta-Leena Metsäpelto [a, *], Jukka Utriainen [b], Anna-Maija Poikkeus [c], Joona Muotka [d], Asko Tolvanen [c], Anu Warinowski [e]

[a] Department of Teacher Education, Alvar Aallon katu 9, P. O. Box 35, FI-40014, University of Jyväskylä, Finland
[b] Finnish Institute for Educational Research, Alvar Aallon katu 9, P. O. Box 35, FI-40014, University of Jyväskylä, Finland
[c] Faculty of Education and Psychology, Alvar Aallon katu 9, P. O. Box 35, FI-40014, University of Jyväskylä, Finland
[d] Department of Psychology, Kärki, Mattilanniemi 6, PO Box 35, FI-40014, University of Jyväskylä, Finland
[e] Faculty of Education, Assistentinkatu 5, 20500, Turku, University of Turku, Finland

HIGHLIGHTS

- Multiple Mini Interview (MMI) format uses many short independent assessments.
- Evidence supporting MMI as reliable tool for initial teacher education selection.
- Applicants and interviewers perceived MMI mostly positively.

ARTICLE INFO

Article history:
Received 2 December 2020
Received in revised form
9 December 2021
Accepted 28 January 2022
Available online 10 February 2022

Keywords:
Multiple mini interviews
Student selection
Initial teacher education
Gender
Age

ABSTRACT

This study investigates the reliability of multiple mini interviews (MMIs) to select students for classroom and special education teacher programs (n = 418) using intraclass correlations and cross-classified multilevel modeling. The results indicated mostly small effects of clustering of applicants to different interviewers and five-station circuits. The largest variance components in the MMI total score were for applicants (63.3%) and measurement error (20.6%), while the variance component for the interviewer was relatively small (11.6–14.4%). The applicants' and interviewers' perceptions were positive. This study provides evidence for the use of MMIs as a reliable tool for initial teacher education selection.

1. Multiple mini interviews as a Selection Tool for Initial Teacher Education Admissions

Some countries have more applicants for university-based initial teacher education (ITE) programs than can be admitted into the limited number of vacancies available. Selection into ITE programs is often based on applicants' general academic achievements in their final year of secondary school (Ingvarson, 2013). In Finland, where this study was conducted, admission into ITE programs is highly competitive (with acceptance rates of around 10%) and involves a broader research-based screening of applicants at entry. The selection phase of ITE and the development of competences over the study years constitute a critical base for building teaching quality, which is a crucial factor in the success of an educational system (Kelly et al., 2018). High teaching quality emerges from a combination of desired preexisting competencies (e.g., strong academic and social skills), effective support for competence growth in teacher education, and continuing professional development throughout teachers' careers (Klassen & Kim, 2017). Recent years have witnessed an increasing interest in

* Corresponding author. Department of Teacher Education, P.O. Box 35, 40014, University of Jyväskylä, Finland.
E-mail addresses: riitta-leena.metsapelto@jyu.fi (R.-L. Metsäpelto), jukka.t.utriainen@jyu.fi (J. Utriainen), anna-maija.poikkeus@jyu.fi (A.-M. Poikkeus), joona.s.muotka@jyu.fi (J. Muotka), asko.j.tolvanen@jyu.fi (A. Tolvanen), anu.warinowski@utu.fi (A. Warinowski).

https://doi.org/10.1016/j.tate.2022.103660

teacher education selection (Bowles et al., 2014; Klassen & Kim, 2018) and a concerted effort to improve the process by developing more reliable and valid selection methods (Klassen & Kim, 2021).

In fields such as medicine, a strong research base for student selection has accumulated, aimed at the development of effective and reliable selection methods (Patterson et al., 2016), but this is not the case in the field of education, where traditional selection methods are considered less than optimal (Klassen & Kim, 2018), and research is scarce. Much of the active research in recent years has focused on developing situational judgment methods for ITE selection, focusing on applicants' abilities to consider situations that teachers might face at work, and judging how they would respond to a potential dilemma using both video-based and text-based tests (Bardach et al., 2021a, 2021b). Despite this important strand of research, available information is scant on the reliability of selection methods in ITE, the extent to which different applicant groups (e.g., males and females or younger and more mature applicants) are treated equally, and on how applicants and interviewers perceive the selection methods. The present study addresses this research gap with respect to one specific student selection approach: multiple mini interviews (MMIs).

1.1. Multiple mini interviews

The MMI format represents a relatively recent development in admissions for ITE and was originally designed to assess applicants' personal and social skills or characteristics in medical education selection (Eva, Rosenfeld, et al., 2004). In MMIs, applicants move through a number of interview stations in which they respond to a set of predefined questions on a topic, dilemma, or case-based scenario while being rated by an interviewer using a standard scoring scheme. Thus, MMIs follow a multiple independent sampling methodology (Hanson et al., 2012), in which an applicant is interviewed successively by several interviewers, each of whom assesses the applicant independently on a specific topic. This format contrasts with semi-structured panel interviews, in which interviewers present open-ended questions that allow a conversation with the interviewee on loosely defined themes. The design of MMIs (e.g., the dimensions being measured, the number and duration of stations, and the scoring system) is adjustable to the specific demands of the institution, which makes it more an assessment approach or format than a standard measurement method (Reiter et al., 2012).

Compared to traditional semi-structured interviews, the MMI format has several advantages. Semi- or unstructured interviews are known to suffer from poor psychometric properties (e.g., Salvatori, 2001; Siu & Reiter, 2009), whereas research on MMIs in medical student selection shows relatively high reliability and low interviewer effects (Knorr & Hissbach, 2014; Patterson et al., 2016; Pau et al., 2013). There is also evidence for criterion-based validity, where MMIs predict academic success in medical education and performance in working life (e.g., medical council examinations, tests of clinical skill performance; see Knorr & Hissbach, 2014; Patterson et al., 2016; Pau et al., 2013; Reiter et al., 2007). MMI scores used in combination with cognitive skill measures have been found to predict a lower likelihood of dropout among psychology students (Makransky et al., 2017), and low MMI scores have been shown to predict academic difficulties (e.g., delayed progression and low course grades) among pharmacy students (Heldenbrand et al., 2016). A strength of this approach is that MMI performance does not appear to benefit from coaching (Griffin et al., 2008) or suffer from violations of MMI test security (Reiter et al., 2006).

The MMI format has consistently been shown to be among the strongest selection methods and is one of the most frequently studied approaches in medical education (e.g., Patterson et al., 2016). Its favorable psychometric properties provide a strong impetus for its application in the field of education. The present study is the first to investigate the reliability and utility of MMIs in student selection for teacher education.

1.2. Reliability of MMIs

Traditional selection interviews have been criticized for being susceptible to biases that stem, for instance, from interviewers' occupational stereotypes and expectations, unstructured interview situations, and flawed judgment based on inferences from limited or biased information (bib_Ebmeier_and_Ng_2005Ebmeier & Ng, 2005). Consequently, an applicant's success in the selection interview may be influenced by the interviewer's characteristics, leading to poorly justified selection decisions. It has been shown that almost 56% of the score variance from traditional interviews for average and low-achieving applicants can be attributed to interviewer variability (Harasym et al., 1996).

An effective means to reduce interviewer bias and increase reliability is a highly structured interview format (e.g., Ebmeier & Ng, 2005) like the multiple mini interview, which is based on predefined dimensions being assessed, uniformly applied standard questions, and systematic scoring rubrics that are employed consistently across all stations and applicants by carefully trained interviewers. The meticulous structuring seeks to reduce and minimize differences between interviewers regarding how they conduct the interview and apply the scoring rubric. Without these precautions, in a situation in which an interviewer has a large number of applicants to assess, there is a risk that assessments will begin to resemble each other and lead to problems of reliability.

Critical preconditions for MMI reliability are that variance due to interviewers or the circuit (i.e., the series of stations) is minimal (Knorr & Hissbach, 2014), and that applicant characteristics explain the majority of variance in MMI scores. Prior research on medical education selection has provided evidence of MMIs' success in reducing “unwanted” variance that is irrelevant for the dimension being measured (Roberts et al., 2010). One example is interviewers' stringency or leniency, which refers to a consistent tendency to award applicants higher or lower scores than is justified by their responses. Some studies have estimated the contribution of different sources of variance to MMI scores and have found that the effect of interviewers' stringency or leniency is relatively small (e.g., 14%; Roberts et al., 2008; see also Yoshimura et al., 2015), and variance due to the circuit is negligible (Hecker & Violato, 2011). The reliability of applicants' scores in MMIs in medical selections has typically been found to be at least marginally satisfactory or good (~0.55−0.8; Dore et al., 2010; Eva, Reiter, et al., 2004; Roberts et al., 2008; Sebok et al., 2014; Uijtdehaage & Parker, 2011; Yoshimura et al., 2015), with a greater number of stations providing higher reliability estimates. Reliability estimates of approximately 0.5 were successfully increased to 0.7 after modifications to MMIs aimed at improving reliability (e.g., replacing an easy station with a more challenging one; Uijtdehaage & Parker, 2011). The scores obtained by applicants at each station often differ, indicating that stations have differing levels of difficulty (Dore et al., 2010; Hecker & Violato, 2011). However, the variation in station difficulty does not generate consistent differences between applicants, as they all go through the same stations. In a case where each station assesses one attribute, the internal consistency of the scores assigned within any one station reflects the degree of measurement error (or lack thereof). Often, however, multiple attributes are measured at a single station, and the internal consistency of such stations can range from low or moderate (Dowell et al., 2012) to high (Dore et al., 2010).

To the best of our knowledge, this study is the first to examine the reliability of MMIs in teacher education selection. The first goal of the present study was to investigate the reliability of MMIs using two approaches: 1) focusing on the effect of applicant clustering to different interviewers, and 2) estimating the contribution of different sources of variance—interviewer, circuit, station, applicant, and measurement error—to the variance of MMI scores.

1.3. Adverse impact of gender and age

Fair treatment of applicants in the selection process is important to ensure equal opportunities for access to ITE. The distribution of male and female students in teacher education programs has been a much-discussed topic (Sabbe & Aelterman, 2007), and an increase in male teachers in primary schools has been called for (Skelton, 2009). In Finland, the teaching profession is more popular with women than men at the primary and secondary levels, a trend common in Western cultures (OECD Education at a Glance, 2019). To ensure equal opportunities for all, it is critical that admission procedures do not favor a particular gender.

Although research has often found MMI scores to be unrelated to gender (Humphrey et al., 2008; Reiter et al., 2012), other research has reported that female applicants for medical school tend to receive higher MMI ratings than male applicants (Barbour & Sandy, 2014; Ross et al., 2017). In addition, research on the impact of age on MMI performance has documented that older applicants outperform younger applicants (Reiter et al., 2012). Therefore, the second goal of the present study was to examine the possible impact of applicant gender and age on MMI performance.

In previous research, it has been suggested that female applicants achieve higher scores, particularly in stations that assess applicants' abilities to understand and share the feelings of another person (empathy), because women are better at these skills than men. The higher achievement of older applicants has been linked to their greater life experience and overall maturity (see Knorr et al., 2019). However, because nonsignificant gender and age effects appear to be the most common finding in medical education selection—including findings from several review articles (Pau et al., 2013; Rees et al., 2016)—we expected our study to replicate this result in ITE selections and not reveal significant gender and age differences.

1.4. Applicant and interviewer reactions

It is important that both applicants and interviewers have confidence in admission procedures and perceive them as fair and valid (McCarthy et al., 2017). Prior research in medical education selection shows that applicants and interviewers generally perceive MMIs positively (Dore et al., 2010; Eva, Reiter, et al., 2004; Patterson et al., 2016; Razack et al., 2009) because of its format of individual interviews and multiple opportunities for the assessment of applicant attributes (Kumar et al., 2009). Although the short duration of interview stations and limited opportunities for applicants to freely discuss their commitment and values have been considered limitations of MMIs (Kumar et al., 2009), applicants have reported feeling that they could demonstrate their communication skills, critical thinking skills, and opinions during interviews (Cox et al., 2015).

Interviewers consider the multistation format better than the traditional panel interview format (Humphrey et al., 2008; Razack et al., 2009), and appreciate the MMI decision-making process because it is free from the possible influence of interviewers on each other, which is typical of panel interviews. In addition, because each applicant is being assessed by several different interviewers, the pressure of assessment is lower and encourages interviewers to use the full scale (Kumar et al., 2009). It is possible that the usability of MMIs is perceived differently in different disciplines; thus, the third goal of the present study was to investigate both applicant and interviewer perceptions of MMIs in teacher education admissions.

1.5. The present study

The present study addresses the following research questions (RQ):

RQ1: What is the reliability of the MMI format in ITE admissions?

- 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 total scores is attributable to the interviewer, circuit, station, applicant, and measurement error?

RQ2: Are there differences in the MMI scores of male and female applicants and between younger and older applicants?

RQ3: What are applicant and interviewer perceptions of MMIs with respect to their validity and usability?

2. Methods

2.1. Participants and procedures

This study focused on the locally contextualized MMIs used in one Finnish teacher education unit to select students for classroom teacher (grades 1–6) and special education teacher programs. Both ITE programs offer a three-year bachelor's degree and a two-year master's degree. The selection procedure included two phases. The first phase (cognitive screening) consisted of a source-based exam (four scholarly articles in the field of education, about 140 pages) with multiple-choice questions measuring conceptual comprehension, ability to recall and connect information correctly, and reasoning. The scores earned in this exam were used to select the top applicants for the second phase, which consisted of an aptitude test, including MMIs. In the present study, we focused solely on the reliability and usability of MMIs as part of the selection process. MMIs had been used in the previous year as an interview format in the second phase of selection. Prior to that, applicants were interviewed for 20 min by two interviewers using a semi-structured format.

2.1.1. Participants

The participants of the present study included applicants seeking admission to classroom teacher and/or special education teacher programs and interviewers who assessed them using the MMI format. Based on the scores earned in the first-phase exam, 482 applicants participated in the second phase. Of these applicants, 418 gave permission for their MMI scores to be used to examine the reliability of MMIs and the differences between the subgroups (RQ1 and RQ2), and 304 additionally agreed to complete a web survey assessing applicant perception of MMIs (RQ3). The applicants were informed about the purposes of the study, and it was emphasized that taking part was voluntary and would not have any influence on admission decisions. All participants gave their written consent to participate. The mean age of the 418 participants was 25.2 ($Mdn = 21.9$, $SD = 7.9$) years, and the majority of the applicants were women (84%).

The MMIs were conducted by the staff of the classroom teacher education and special education units and by in-service teachers at

the university teacher training school ($n = 53$). Each interviewer assessed 15 to 88 applicants. Of the interviewers, 28 (53%) participated in a web survey on their perceptions of the MMIs. Of the respondents, 75% ($n = 21$) were women, 25% ($n = 8$) were men, and their mean age was 48.9 years ($SD = 10.8$). The interviewers responding to the survey gave their written consent for participation.

2.1.2. Procedure

Multiple Mini Interviews. The MMI circuit had five stations, each lasting 5 min, with a 3-min turnaround between the stations. The 5-min duration has been found to be a cost-effective solution with only a minimal reduction in reliability compared to an 8-min duration (Dodson et al., 2009). MMIs have been found to generate reliable interview results using only five stations (Fraga et al., 2013), and MMIs with a small number of stations are not uncommon (see Klassen & Kim, 2021, for implementation of three-station MMIs in the UK). Each applicant rotated through the five-station circuit, meeting a single interviewer at each station. Five simultaneous five-station circuits were operated over four consecutive days, totaling 20 circuits. Each interviewer was involved in the MMIs on one to four days. To make effective use of resources, interviewers did not remain in one circuit, but conducted interviews in several circuits over the four days. Of the 53 interviewers, 18 interviewed at only one circuit, 34 at two different circuits, and 1 at four circuits. The interviewers switched between circuits, but always remained at the same station. All interviewers received 4 h of training consisting of the general aims and implementation of MMIs and extensive training on the administration and scoring of their respective stations.

The MMI stations and criterion-referenced scoring schemes were highly structured in terms of interview questions and the evaluation of responses to ensure that the MMIs were administered consistently to all applicants. The same questions were asked of each applicant at the respective stations, and the rating scales for scoring applicants' responses were anchored with descriptions and examples of scores. The scoring rubric required interviewers to assess applicants against a set of predefined criteria without reference to the achievements of other applicants. The stations included combinations of different task contents (see Table 1): *situation-based content* (applicants were presented with a scenario requiring them to imagine and describe what they would do if they were to encounter a particular problem), *experience-based content* (applicants were required to recall their particular experiences and the behaviors they demonstrated), *performance content* (applicants were required to use the skill being assessed to solve a task or a problem), and *reflection* (applicants were required to consider and reflect on some subject matter or idea) (e.g., Eva & Macala, 2014). At each station, an applicant's performance was assessed on a rating scale from 0 to 12, which was based on the total of scores assigned for station-specific subscales (e.g., the total score was calculated by adding the scores from three subscales, each with a four-point maximum; see Table 5). In the statistical analyses, we used the applicant's MMI total score, which was calculated as a mean of MMI station scores, as well as subscale scores assigned to applicants within each station.

An admission committee consisting of 10 senior staff members designed the MMI stations. This work was guided by a national initiative of seven universities to improve and unify the student selection processes for ITE in Finland (Student Selection to Teacher Education in Finland—Anticipatory Work for Future; research project funded by the Finnish Ministry of Education and Culture). It included the process of constructing a teacher competence model specifying the key competence domains perceived to be critical for the teaching profession in the Finnish educational landscape (Metsäpelto et al., 2021). In this work, high-quality teaching was characterized as learner-centered and constructivist, with an emphasis on teaching interactions, students' active learning and problem solving, and teachers' emotional and learning support for a diverse student body. Station development was based on the selected set of attributes that were considered to form the basis for developing these skills in the context of teaching and learning and indicating applicants' general suitability for the teaching profession. Following Eva, Reiter, et al. (2004), applicants were not expected to have specialized knowledge or show expertise in teaching.

Applicant and Interviewer Reactions. Upon completion of the MMIs and before leaving the test site, applicants were invited to complete a web survey regarding their perceptions of the MMIs. The web survey was administered separately from the selection procedure, and applicants were informed that responses to the survey would not have any bearing on the selection itself. Interviewers were approached by e-mail approximately six weeks after the MMIs to ask for their participation in a web survey on their perceptions of the MMI and their evaluation of its usability.

Perceptions of the MMIs were collected from the applicants and interviewers using an identical nine-item questionnaire (Chan et al., 1998). The participants were asked to evaluate the nine statements using a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree). The questionnaire included three subscales, each with three items: 1) fairness (e.g., I feel that using MMIs to select applicants for the teacher education programs is fair); 2) perceived predictive validity (e.g., The results from the MMIs can predict how well an applicant will perform in teaching work); and 3) face validity (e.g., The actual content of the MMIs are related to teaching work). Cronbach's alphas for the scales were calculated as indicators of scale reliability. The alphas for applicants' and interviewers' ratings were as follows (interviewer alphas in parentheses): fairness = 0.55 (0.66), predictive validity = .63 (0.73), and face validity = 0.65 (0.87).

The web survey for interviewers also included an additional 10 items about the feasibility of MMIs (e.g., The MMI was easier to carry out than the previously used [semi-structured panel] interview method). The interviewers evaluated the items using a 5-

Table 1
MMI Stations, Format, and Content Type.

Station	Main target of assessment	Format	Content type
1	Social skills in perspective taking, i.e., understanding another person's thoughts and feelings	Scenario and questions	Situation-based, performance
2	Cultural competence, i.e., skills in relating to cultural diversity	Interview questions	Reflection
3	Motivation for pursuit of a teaching career	Interview questions	Reflection
4	Skills in managing emotions	Scenario and questions	Situation-based, experience-based
5	Collaboration skills, i.e., skills for problem solving in a teamwork situation	Problem-solving task requiring collaboration between the applicant and the interviewer	Performance, reflection

point Likert scale (1 = strongly disagree; 5 = strongly agree). In the analyses, these were used as single items to describe the interviewers' perceptions regarding the usability and applicability of MMIs in ITE selection.

2.2. Analysis strategy

We first present descriptive statistics to provide an overview of applicant scores in the five MMI stations and the MMI total score, and the Pearson correlations between them. The five MMI station scores were normally distributed (skewness values ranging from −0.82 to 0.23 and kurtosis values ranging from −0.84 to 0.60), allowing the use of parametric statistical analyses.

Second, intraclass correlations (ICCs) were calculated to investigate 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 five-station circuits (i.e., circuit effect) (RQ1). The ICC is a measure of the relatedness of observations within a cluster, and it ranges from 0 to 1 (Killip et al., 2004). An ICC of 0 indicates that there is no correlation of observations within a cluster, and when an ICC is 1, all observations within a cluster are identical. In the present study, as an index of reliability, we examined whether the MMI scores of an individual interviewer were more similar than the scores derived from different interviewers.

Third, to analyze the complex structure of variation in the MMI total scores in more detail, we used a cross-classified multilevel model. This model is based on generalizability theory, which estimates multiple sources of measurement error with the aim of designing measurement procedures that minimize error (Brennan, 2010; Webb et al., 2006). It estimates the contributions of different factors (i.e., variance components) to the variance of the MMI total score. We calculated a two-level cross-classified multilevel model with three factors to estimate the variance components for the interviewer and circuit to which the applicant was assigned. We also examined the variance component of the station, investigating the mean differences in scores obtained by applicants at each station (RQ1). Cross-classification is applied in situations where the data hierarchy is ambiguous (Hox, 2010; Rasbash & Browne, 2008). For example, students are members of schools and neighborhoods, but not all students from the same neighborhood go to the same schools; hence, neighborhoods and schools are crossed, while students are nested within them (Hox, 2010; Rashbash & Browne, 2008). When the contributing factors for the total variation of assessment are crossed, cross-classified multilevel modeling allows a statistical model to be built that takes into account the effects of separate variance components on the outcome variable (e.g., Marsh et al., 2008).

In the present study, the outcome was the scores earned at each of the five stations. The two-level cross-classified multilevel modeling aimed at exploring the extent of variance attributable to the interviewer, circuit, and station. The mathematical formula for calculating the cross-classified model is presented in Fig. 1. Applicants moved through five stations; hence, the stations were not crossed, and they were treated as a fixed effect in the cross-classified model. The effects of the stations' relative difficulty level were calculated by using dummy variables (e.g., station 1; 0 = no, 1 = yes), and were regressed on the applicants' scores at each station.

Applicants were, however, crossed with interviewers, as five interviewers out of the total pool of 53 assessed each applicant. In addition, applicants were crossed with circuits because they attended one of the 20 circuits. Not all interviewers remained in one circuit, but some conducted interviews in two or even four different circuits over the four admission exam days. The remaining variance in the cross-classified model was residual variance, which could not be explained by the variables in the model. The calculation of ICCs and the cross-classified multilevel models was accomplished using Mplus 7.4 (Muthén & Muthén, 2015). The confidence intervals for the ICCs were calculated by simulation, where sampling variances (between and within levels) were used.

In the cross-classified modeling, the residual variance had two sources of variation: applicants' true score variation and measurement error. However, the cross-classified multilevel modeling did not allow us to separate out these sources of variance. As this study aimed to investigate the degree to which MMIs detected valid systematic differences between applicants (RQ1), we next aimed to calculate the variance attributable to applicants. Recall that each station assessed one personal or social skill using two to six subscales. Using subscale scores, we calculated Cronbach's alphas for each station. Cronbach's alpha is a measure of scale reliability (internal consistency), and it provided an estimate of how much of the variation in each station was explained by applicants' true scores and how much was measurement error. The use of Cronbach's alpha to partition applicants' true score variance and measurement error variance is based on classical test theory (Brennan, 2010; Webb et al., 2006) and the formula $X = T + E$, where X, T, and E are observed, true, and error score random variables, respectively (Brennan, 2010). It follows from the mathematical formula that once the measurement error (E) is defined, the true score is unambiguously derived. The true score reflects the stable or

Within level

$\text{MMITS}_{ijk} = \nu_j + \gamma_k + s_1 + s_2 + s_3 + s_4 + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \sim \text{N}(0, \delta_\varepsilon^2)$

$i = 1, 2, .., 418$ *individuals*

s_x *is* 1 if station number is x, else the value is 0 (the fifth station is set as a reference)

Between interviewer level

$\nu_j \qquad \nu_j \sim \text{N}(0, \delta_\upsilon^2) \qquad j = 1, 2, .., 53$ *identification number of the interviewer*

Between circuit level

$\gamma_k \qquad \gamma_k \sim \text{N}(\mu, \delta_\gamma^2) \qquad k = 1, 2, .., 20$ *identification number of the circuit*

Fig. 1. *Multiple Mini Interview Total Scores (MMITS) Two-level Cross-classified Model. Note.* At within-level, the MMITS consisted of the scores that applicants earned at the five stations. The variance of MMITS summed up differences between stations ($s\_1 + s\_2 + s\_3 + s\_4$), interviewers $\delta\_\upsilon\hat{2}$, circuits $\delta\_\gamma\hat{2}$ and residual $\delta\_\varepsilon\hat{2}$. Residuals have two sources of variation, individuals true score variation and measurement error.

Table 2
MMI *Station Score Means for Total Sample and For the Male and Female Applicants, and Correlations Between Study Variables.*

Score/variable	All applicants		Female		Male		df	*t*	*p*	Correlations					
	M	*SD*	*M*	*SD*	*M*	*SD*				1	2	3	4	5	6
1. Station 1	7.5	2.1	7.6	2.0	7.1	2.1	416	1.70	.089	–					
2. Station 2	9.0	2.0	9.0	2.0	8.8	2.1	416	0.94	.349	.18**	–				
3. Station 3	8.3	2.0	8.3	1.9	8.3	2.2	416	0.04	.969	.04	.33**	–			
4. Station 4	9.7	1.9	9.7	1.9	9.6	2.1	416	0.29	.770	.19**	.09	.10*	–		
5. Station 5	6.8	2.1	6.7	2.0	7.3	2.4	81.93	−1.85	.068	.15**	.17**	.17**	.12*	–	
6. MMI Total score	8.2	1.1	8.3	1.1	8.2	1.4	416	0.28	.783	.55**	.62**	.57**	.51**	.58**	–
7. Age	25.2	7.9	25.5	8.5	23.6	3.7	–	–	–	-.01	-.03	.00	-.01	.10*	-.01

Note. The scores ranged between 0 and 12. *$p < .05$ **$p < .01$. 1 Social skills, 2 Cultural competence 3 Teacher motivation, 4 Emotion management, 5 Collaboration skills.

nonrandom individual differences between applicants.

As is generally known, the measurement errors of the stations do not correlate, so they can be summed. The sum of the station-specific measurement errors was then used as an overall indicator of the amount of measurement error in the MMI total score. The total estimated measurement error was subtracted from the residual estimated in the cross-classified model, and the resulting estimate indicated the true score variance for the applicants.

Fourth, an independent sample *t*-test was used to compare the scores earned at the five MMI stations and the MMI total score between male and female applicants. Pearson correlation coefficients were calculated between applicant age and MMI station scores and total scores (RQ2). Finally, maximum and minimum scores, mean scores, and standard deviations were used to describe how the applicants and interviewers perceived the MMIs (RQ3).

3. Results

3.1. Descriptive statistics

Table 2 shows that there was a relatively large variation in the mean scores between the stations. Applicants received high scores, particularly in Station 4 (emotion management; $M = 9.7$), while their scores in Station 5 were, on average, the lowest (collaboration skills; $M = 6.8$). Correlations between the stations ranged from low to moderate ($R = 0.04–0.33$; the highest correlation was between cultural competence and teacher motivation), indicating that the stations measured separate personal or social skills.

3.2. Interviewer and circuit effects

The analysis of intraclass correlations provided information on the extent to which variation in applicant scores at each station was explained by the clustering effect of the interviewer and the circuit. The findings presented in Table 3 show that three stations had ICCs of less than 0.10: Stations 1, 2, and 5, assessing applicants' social skills, cultural competence, and collaboration skills. Thus, nearly all the measured variance in these stations was attributable to sources of variance other than factors related to the interviewer or circuit. In two stations, however—Station 4 (emotion management) and Station 3 (motivation for teaching career)—ICCs were higher, 0.18 and 0.28 respectively, indicating that scores of applicants having the same interviewer resembled each other more strongly in these two stations than in other stations. The results for analyses with the circuit as a clustering variable showed that the five-station circuit to which the applicant had been assigned had a very small contribution to the variance in the total MMI scores.

Table 3
I*ntraclass Correlations (ICC) of the Five MMI Stations (in Ascending Order of the ICC) and the Five-station Circuits.*

Clustering variable	ICC	95% Confidence interval	
		Lower bound	Upper bound
Interviewer [a]			
Station 1	.05**	.00	.11
Station 5	.06*	.00	.12
Station 2	.08*	.00	.16
Station 4	.18**	.03	.32
Station 3	.28**	.11	.46
5-station circuit			
Total MMI-score	.07*	.00	.14

Note: * < 0.05, ** < 0.01. [a] 1 Social skills, 2 Cultural competence 3 Teacher motivation, 4 Emotion management, 5 Collaboration skills.

3.3. Variance components of the MMI total scores

The findings of the cross-classified multilevel modeling, delineating the variance components of the MMI total score, are shown in Table 4. All variance components were statistically significant. We first estimated all sources of variance that explained the applicant's MMI total score (interviewer, circuit, and station), and the remaining residual variance was considered an aggregate of variance attributable to applicants and measurement error. The variance in the MMI total score explained by the interviewer was 11.6%, whereas the variance explained by the circuit was minimal (1.4%). This means that there were relatively small differences in the MMI total scores as a function of the interviewer or the specific circuit to which the applicant was assigned.

The variance related to the station was slightly larger (19.7% of total variance) and indicated that a significant portion of the variance in the MMI total scores was due to differences between stations, that is, the relative difficulty of each station. The residual variance component was large and statistically significant (67.4%) and included differences between applicants in the dimensions assessed, as well as measurement error variance. Table 4 also shows the estimated sources of variance that explain applicant-to-applicant variations in the MMI total score. Because all applicants moved through the same stations, the relative difficulty of the stations exerted a uniform effect on all applicants and did not generate variations between applicants. When we removed the station effect from the sources of variation in the MMI total score, the estimated variance components included the interviewer (14.4%), circuit (1.7%), and residual (83.9%).

To differentiate between the true score variance related to applicants and the measurement error variance, we calculated the total measurement error in the five stations. This was accomplished

Table 4
Variance Components for Multiple Mini Interview Total Scores.

Source of variance	MMI total score		Applicant-to-applicant variations in MMI total score	
	Variance component	Percentage of total variance	Variance component	Percentage of total variance
Interviewer	0.60	11.6	0.60	14.4
Circuit	0.07	1.4	0.07	1.7
Station	1.02	19.7	–	–
Residual	3.49	67.4	3.49	83.9
Total	5.18	100	4.16	100

Table 5
Cron*bach Alpha Reliabilities, Variances, and Variances of Measurement Error.*

Station	Nr of subscales at a station	*n*	Variance	Cronbach alpha reliability	Variances of measurement error
Station 1	3	387	4,209	0,77	0,968
Station 2	4	390	3,889	0,77	0,894
Station 3	6	412	3,878	0,53	1,823
Station 4	3	316	3,663	0,64	1,319
Station 5	2	369	4,325	0,62	1,644
Total	18	–	–	–	6,647

1 Social skills, 2 Cultural competence 3 Teacher motivation, 4 Emotion management, 5 Collaboration skills.

by estimating the measurement error in each station using information about subscale reliability provided by Cronbach's alpha analysis. As shown in Table 5, the reliability of the stations ranged from 0.53 (Station 3) to 0.77 (Stations 1 and 2). We summed the variances of measurement error in each station, which resulted in a total measurement error variance of 6.647, while the MMI total score variance—based on the MMI summary score—was 32.157. The proportion of total measurement error variance from the total score variance was 20.6%. The total measurement error variance was then subtracted from the variance component of the residual variance (83.9%; see Table 4). The resulting figure represents the applicants' true score variance, which accounted for 63.3% of the total variation in the MMI.

3.4. Gender and age differences

There were no significant differences in the MMI scores between male and female applicants, as shown in Table 2. Furthermore, the relationship between age and MMI scores showed no significant correlations, apart from Station 5 (collaboration skills). Applicants who were older were evaluated as having better collaboration skills, although this association was weak ($r = 0.10$).

3.5. Interviewer and applicant reactions

Analysis of the ratings of both applicants and interviewers indicated that MMIs were, on average, perceived to be fair and to have high face validity (Table 6). The survey responses indicated that MMIs offered equal opportunities for all applicants to contend for a place in the ITE program, and that the contents of the MMIs, with respect to the skills they assessed, were connected to the work of teachers. The ratings concerning perceived predictive validity were somewhat lower, suggesting that applicants were less certain about the MMIs' ability to predict an applicant's subsequent performance as a teacher. The examination of maximum and minimum scores and standard deviations indicated that scores were spread out over a wide range of values, suggesting relatively large differences within both applicants' and interviewers' perceptions of fairness and the face and predictive validity of MMI.

The interviewers' ratings of MMIs as a feasible assessment format for ITE selection were quite positive. MMIs were considered highly suitable for use as an entrance exam for ITE ($M = 4.39$). The interviewers' ratings indicated that the instructions to implement the MMIs (e.g., the structured format of stations with detailed questions, timing, and scoring) were clear ($M = 4.14$), and MMIs

Table 6
Applicants' a*nd Interviewers' Responses to Statements on MMIs Scored on a 5-point Likert Scale (1–5).*

Item/Scale	Minimum score	Maximum score	Mean score	*SD*
Applicants				
Face validity	2.33	5.00	4.01	0.58
Fairness	2.00	5.00	3.92	0.57
Predictive validity	1.00	4.67	2.66	0.61
Interviewers				
Face validity	2.00	5.00	3.89	0.79
Fairness	2.33	4.67	3.77	0.70
Predictive validity	1.67	4.33	2.90	0.61
The MMI is a suitable method of selection as part of the ITE selection process.	3.00	5.00	4.39	0.74
The length of the MMI stations (5 min) was long enough.	2.00	5.00	4.18	0.86
The instructions I received to implement the MMI were clear and understandable.	2.00	5.00	4.14	0.71
The level of difficulty in MMI stations was appropriate for those applying for ITE.	2.00	5.00	3.96	0.79
For me it was easier to conduct the MMIs than the interview format we earlier had.	1.00	5.00	3.86	1.15
The competence or skill I was assessing was relevant and easy to understand.	2.00	5.00	3.79	0.83
The amount of time allotted for rating each applicant (3 min) was sufficient.	1.00	5.00	3.79	1.32
The training for the interviewers was adequate.	1.00	5.00	3.61	1.07
The contents of the MMI stations were clear and easy to understand for applicants.	2.00	5.00	3.57	1.03
The workload involved in preparing for the MMIs was excessive.	1.00	5.00	2.32	0.77

were easier to implement than the previously used semi-structured panel interview format ($M = 3.89$). The interviewers also evaluated the time allotted for each station as sufficiently long in duration ($M = 3.79$). It is noteworthy, however, that, on average, the interviewers were less satisfied with the training for administering the MMI stations and the degree of clarity and comprehensibility of content for the applicants. This finding indicates the need to improve these aspects in future ITE selection.

4. Discussion

To the best of our knowledge, the present study is the first to investigate the reliability and utility of MMIs for student selection in teacher education. The analysis of the ICCs showed that the effect of the clustering of applicants to different interviewers and circuits was mostly small (under 10%). In two stations, however, the clustering effects were higher (0.18 and 0.28), indicating an elevated interviewer effect. The findings of cross-classified multilevel modeling indicated that the variance in the MMI total score explained by the interviewer or the circuit was relatively low or minimal, while the variance component for the station was somewhat higher (19.7% of total variance), indicating varying levels of difficulty between stations. The measurement error accounted for 20.6% of variance, reflecting challenges in establishing the internal consistency of the scores assigned within any one station. An estimated 63.3% of the variance between the scores could be attributed to the applicant, reflecting the marginally satisfactory reliability of the applicants' scores in the MMIs. The findings further showed that the associations of MMI scores with applicants' gender or age were minimal or nonexistent. The perceptions of the applicants and interviewers of MMIs, as indicated by ratings in a web survey, were mostly positive. Taken together, the present study demonstrates that the MMI is a feasible selection tool with satisfactory reliability for a high-stakes entrance examination, determining who will be the most suitable applicants to enter teacher education programs.

The present analyses show that the clustering of applicants to different interviewers explained only a small amount of variance ($\leq$ 8%) for three out of five MMI stations. This means that scores assigned within the pool of applicants of the same interviewer did not resemble each other more than scores of other applicants by other interviewers; thus, the treatment of applicants in most stations was reliable and consistent. In the other two stations, intraclass correlations were somewhat higher, suggesting that interviewer bias may have partly affected selection scores, even in highly structured tasks with uniform, criterion-based scoring rubrics. More research is needed to understand the various sources of interviewer bias in MMI stations to improve assessment reliability. It should be noted that even though the interviewer effect based on ICCs was found to be slightly higher in the two stations—notably the station assessing applicants' motivations for pursuing a teaching career—the effect was diluted when the ICC (0.07) of the five-station circuit was taken into account. This means that the reliability of the overall assessment of applicants across the five stations was within acceptable limits.

Further evidence of the reliability of MMIs was obtained from cross-classified multilevel modeling. When the sources of variance in the MMI total score were examined, it was found that the variance component attributable to the interviewer was only slightly above 10%. In previous studies, the amount of variance attributable to interviewers' stringency or leniency has also been relatively small (Roberts et al., 2008; Yoshimura et al., 2015). Similarly, the variance due to the circuit has been found to be negligible (Hecker & Violato, 2011). The findings of the cross-classified modeling showed that the effect of the station was about a fifth of the overall variance of the MMI total score, suggesting that the level of difficulty between the stations varied significantly (i.e., the applicants, on average, were assigned lower scores on some stations than others). This result is not surprising, as the stations were designed to function independently, and their level of difficulty was not calibrated to other stations. The differences in station difficulty do not compromise the reliability of the method because all applicants go through the same stations. The variance component of the station, that is, the difficulty of the task, was twice as large as the variance component of the interviewer, which further supports the conclusion that interviewers generally performed consistently in administering and scoring the MMIs.

The reliability of the applicants' MMI score (0.633) is comparable with the previously reported range for MMIs (~0.55–0.8; see Dore et al., 2010; Eva, Reiter, et al., 2004; Roberts et al., 2008; Sebok et al., 2014; Uijtdehaage & Parker, 2011; Yoshimura et al., 2015), but on a sample of ITE applicants whose MMI performance has not been previously examined. This reliability (although marginally acceptable) was a positive finding, especially since the MMI procedure was implemented cost-effectively, using only five 5-min stations. Unlike most other studies, this study also examined the reliability of two or more subscales to measure each attribute. We found that there were large variations between stations in the internal consistency of the measurements, with Cronbach's alpha ranging between 0.53 and 0.77. Moreover, approximately 21% of the total variance in the MMI scores was explained by measurement error, although stations were highly structured, and the scoring rubric was criterion-based. More research attention should be directed to investigating the reliability of assessments within stations and developing ways to increase their internal consistency.

Taken together, the analysis of ICCs and the variance components of the MMI total scores indicated acceptable reliability for the MMIs used in the ITE selections. Thus, it can be concluded that the MMI format is effective in reducing the unreliability that has long been associated with selection interviews for teacher education programs. The content for selection, highly structured interview format, and detailed uniform scoring rubric are likely explanations for these findings. Hence, the use of the MMI format in ITE student selections can be seen as successfully responding to the call for a more carefully defined and reliable evaluation process.

The results further show that the MMI ratings given to men and women were similar and did not favor either gender. This result runs counter to prior findings in other fields, such as those in which female applicants to medical schools have been found to outperform male applicants in MMIs (e.g., Barbour & Sandy, 2014; Ross et al., 2017). We also found that the age of applicants was only marginally related to performance in MMIs and was significant for only one station: applicants' collaboration skills. Examination of differences between certain subgroups (e.g., based on gender or age) has often been neglected in student selection studies (Klassen & Kim, 2018), and the results of this study provide valuable information on this issue with respect to ITE selection. However, further studies are needed to examine the possible adverse impact on performance in MMIs for applicants belonging, for instance, to different language groups and ethnic minorities.

Successful student selection in any educational institution is greatly strengthened if both the applicants and members of the selecting institute perceive the admissions procedure as fair and well justified. In this study, analyses of the ratings of both applicants and interviewers indicated that they perceived the MMI to be fair, and it was found to have fairly high face validity. These results are in line with earlier findings documenting positive reactions toward MMIs in medical education selection (Dore et al., 2010; Eva, Reiter, et al., 2004; Patterson et al., 2016; Razack et al., 2009). It should be noted, however, that the applicants and interviewers

rated the predictive value of MMI with respect to subsequent success as a teacher less positively. This might be explained by the multidimensional nature of teacher competences, which comprise a large set of skills that may be difficult to define and capture in one admission interview. Admission interviews can target only a relatively narrow area of the applicants' competences; hence, both applicants and interviewers may rightfully feel that a broader assessment is needed to accurately predict future success in the teaching profession. In future studies, investigating the true predictive validity of MMIs in the field of education using longitudinal follow-up data would be highly valuable.

The web survey results were promising in terms of the future development of student selection, as the interviewers' ratings clearly indicated that the MMI format was perceived to be easier to implement than the previously used traditional panel interview format. These results corroborate prior findings, which have similarly reported that interviewers consider MMIs to be superior over semi-structured interviews and easier to use (Eva, Rosenfeld, et al., 2004; Humphrey et al., 2008; Razack et al., 2009). However, there was some variation in the interviewers' ratings regarding the ease of use of MMIs, which suggests that some members of the interviewer pool may need more training and clarification of instructions to improve their confidence. Taking into account both applicants' and interviewers' positive perceptions of MMIs, it can be concluded that introducing MMIs to the ITE selections was seen as a welcome improvement in the validity and reliability of the admission interviews.

4.1. Limitations

One limitation of the present study is that, due to the MMIs being conducted in a high-stakes situation in actual student selections, certain elements regarding the reliability of the MMI could not be investigated. For example, it was not possible to use more than one interviewer per station because the pool of interviewers could not be increased, and budgetary and time constraints did not allow an increase in the number of assessment days. In light of prior studies, the use of five stations in the present MMIs was the minimum recommended in research on the effect of the number of stations on the reliability of MMIs (Fraga et al., 2013). Despite these challenges, the effects of the interviewer, circuit, and station were low or moderate for each individual station, and the reliability of the applicants' scores in the MMIs was satisfactory, corroborating prior findings.

Since we did not collect information about the interviewers' work experience, education, personality characteristics, attitudes, or other individual attributes, it was not possible to examine factors that could explain variability in the MMI ratings between the interviewers. The generalizability of the results is limited because only one educational institution participated in this study. While the results need to be replicated in future studies, including several teacher education institutions, they align well with earlier findings that have shown the acceptable reliability of MMIs in medical education selections.

5. Conclusions

This study reported findings from the successful adoption of MMIs in teacher education admission interviews. A highly structured interview format has been suggested to function as a remedy for interviewer bias (Ebmeier & Ng, 2005), and the low interviewer effects and satisfactory reliability of the MMIs to assess applicants' personal and social skills noticed in this study corroborated this view. However, reliability coefficients at 0.80 or above have been considered sufficiently reliable, especially if assessment tools are used to make decisions that have significant consequences (Webb et al., 2006). Thus, MMIs must be further developed to achieve higher reliability. In the present study, the MMI format was well received by both applicants and interviewers, which is a key issue when developing new methods for student selection. In light of this study, the reliability of ITE student selections could be improved by adopting a more rigorous approach to the development of selection procedures by responding to the call for research-based admissions (Thomson et al., 2011) and by utilizing selection methods that have been demonstrated to work in other fields, such as medical education. Future studies on the use of MMIs in teacher selection should focus on the power of MMIs in predicting study success in teacher education and, ultimately, teaching quality after the transition to working life.

Credit author statement

Riitta-Leena Metsäpelto: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition; Jukka Utriainen: Data curation, Methodology, Formal analysis, Writing – original draft; Anna-Maija Poikkeus: Conceptualization, Writing – original draft; Joona Muotka: Methodology, Formal analysis; Asko Tolvanen: Methodology, Formal analysis, Writing – review & editing; Anu Warinowski: Funding acquisition, Writing – original draft.

Funding

This study has been financed by the Finnish Ministry of Education and Culture (Nr. OKM/47/523/2017) and Academy of Finland (Nr. 292466; 342191)

Declaration of competing interest

The authors declare that there is no conflict of interest.

References

Barbour, M. E., & Sandy, J. R. (2014). Multiple mini-interviews for selection of dental students: Influence of gender and starting station. *Journal of Dental Education, 78*(4), 589–596. https://doi.org/10.1002/j.0022-0337.2014.78.4.tb05710.x

Bardach, L., Rushby, J. V., Kim, L. E., & Klassen, R. M. (2021). Using video-and text-based situational judgement tests for teacher selection: A quasi-experiment exploring the relations between test format, subgroup differences, and applicant reactions. *European Journal of Work & Organizational Psychology, 30*(2), 251–264. https://doi.org/10.1080/1359432X.2020.1736619

Bardach, L., Rushby, J. V., & Klassen, R. M. (2021). The selection gap in teacher education: Adverse effects of ethnicity, gender, and socio-economic status on situational judgement test performance. *British Journal of Educational Psychology, 91*, 1015–1034. https://doi.org/10.1111/bjep.12405

Bowles, T., Hattie, J., Dinham, S., Scull, J., & Clinton, J. (2014). Proposing a comprehensive model for identifying teaching candidates. *Australian Educational Researcher, 41*, 365–380. https://doi.org/10.1007/s13384-014-0146-z

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471–485. https://doi.org/10.1037/0021-9010.83.3.471

Cox, W. C., McLaughlin, J. E., Singer, D., Lewis, M., & Dinkins, M. M. (2015). Development and assessment of the multiple mini-interview in a school of pharmacy admissions model. *American Journal of Pharmaceutical Education, 79*(4), 53. https://doi.org/10.5688/ajpe79453

Dodson, M., Crotty, B., Prideaux, D., Carne, R., Ward, A., & De Leeuw, E. (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. 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), 60–63. https://doi.org/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.
Ebmeier, H., & Ng, J. (2005). Development and field test of an employment selection instrument for teachers in urban school districts. *Journal of Personnel Evaluation in Education, 18*(3), 201–218. https://doi.org/10.1007/s11092-006-9021-4
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. https://doi.org/10.1111/medu.12402
Eva, K. W., Reiter, H. I., Rosenfeld, J., & Norman, G. R. (2004). The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Academic Medicine, 79*(6), 602–609.
Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (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
Fraga, J. D., Oluwasanjo, A., Wasser, T., Donato, A., & Alweis, R. (2013). Reliability and acceptability of a five-station multiple mini-interview model for residency program recruitment. *Journal of Community Hospital Internal Medicine Perspectives, 3*(3–4), 21362. https://doi.org/10.3402/jchimp.v3i3-4.21362
Griffin, B., Harding, D. W., Wilson, I. G., & Yeomans, N. D. (2008). Does practice make perfect? The effect of coaching and retesting on selection tests used for admission to an Australian medical school. *Medical Journal of Australia, 189*(5), 270–273. https://doi.org/10.5694/j.1326-5377.2008.tb02024.x
Hanson, M. D., Kulasegaram, K. M., Woods, N. N., Fechtig, L., & Anderson, G. (2012). Modified personal interviews: Resurrecting reliable personal interviews for admissions? *Academic Medicine, 87*(10), 1330–1334. https://doi.org/10.1097/ACM.0b013e318267630f
Harasym, P. H., Woloschuk, W., Mandin, H., & Brundin-Mather, R. (1996). Reliability and validity of interviewers' judgments of medical school candidates. *Academic Medicine, 71*(1), 40–42.
Hecker, K., & Violato, C. (2011). A generalizability analysis of a veterinary school multiple mini-interview: Effect of number of interviewers, type of interviewers, and number of stations. *Teaching and Learning in Medicine, 23*(4), 331–336. https://doi.org/10.1080/10401334.2011.611769
Heldenbrand, S. D., Flowers, S. K., Bordelon, B. J., Gubbins, P. O., O'Brien, C., Stowe, C. D., & Martin, B. C. (2016). Multiple Mini-Interview performance predicts academic difficulty in the PharmD Curriculum. *American Journal of Pharmaceutical Education, 80*(2), 1–8. https://doi.org/10.5688/ajpe80227
Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.
Humphrey, S., Dowson, S., Wall, D., Diwakar, V., & Goodyear, H. M. (2008). Multiple mini-interviews: Opinions of candidates and interviewers. *Medical Education, 42*(2), 207–213. https://doi.org/10.1111/j.1365-2923.2007.02972.x
Ingvarson, L. (2013). Recruitment and selection in teacher education. In L. Ingvarson, J. Schwille, M. T. Tatto, G. Rowley, R. Peck, & S. L. Senk (Eds.), *An analysis of teacher education context, structure, and quality-assurance arrangements in TEDS-M countries: Findings from the IEA teacher education and development study in mathematics (TEDS-M)* (pp. 165–209) (2013) https://files.eric.ed.gov/fulltext/ED545244.pdf#page=166.
Kelly, S., Pogodzinski, B., & Zhang, Y. (2018). Teaching quality. In B. Schneider (Ed.), *Handbook of the sociology of education in the 21st century* (pp. 275–296). Springer. https://doi.org/10.1007/978-3-319-76694-2_12.
Killip, S., Mahfoud, Z., & Pearce, K. (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. M., & Kim, L. E. (2017). *Assessing critical attributes of prospective teachers: Implications for selection into initial teacher education programs*. http://eprints.whiterose.ac.uk/124823/.
Klassen, R. M., & Kim, L. E. (2018). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review, 26*, 32–51. https://doi.org/10.1016/j.edurev.2018.12.003
Klassen, R. M., & Kim, L. E. (2021). *Teacher selection: Evidence-based practices*. Springer.
Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: Same concept, different approaches. *Medical Education, 48*(12), 1157–1175. https://doi.org/10.1111/medu.12535
Knorr, M., Meyer, H., Sehner, S., Hampe, W., & Zimmermann, S. (2019). Exploring sociodemographic subgroup differences in multiple mini-interview (MMI) performance based on MMI station type and the implications for the predictive fairness of the Hamburg MMI. *BMC Medical Education, 19*(1), 1–12. https://doi.org/10.1186/s12909-019-1674-z
Kumar, K., Roberts, C., Rothnie, I., Du Fresne, C., & Walton, M. (2009). Experiences of the multiple mini-interview: A qualitative analysis. *Medical Education, 43*(4), 360–367. https://doi.org/10.1111/j.1365-2923.2009.03291.x
Makransky, G., Havmose, P., Vang, M. L., Andersen, T. E., & Nielsen, T. (2017). The predictive validity of using admissions testing and multiple mini-interviews in undergraduate university admissions. *Higher Education Research and Development, 36*(5), 1003–1016. https://doi.org/10.1080/07294360.2016.1263832
Marsh, H. W., Martin, A. J., & Cheng, J. H. S. (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
McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing "so what?,""What's new?," and "where to next?". *Journal of Management, 43*(6), 1693–1725. https://doi.org/10.1177/0149206316681846
Metsäpelto, R.-L., Poikkeus, A.-M., Heikkilä, M., Husu, J., Laine, A., Lappalainen, K., Lähteenmäki, M., Mikkilä-Erdmann, M., & Warinowski, A. (2021). *A multidimensional adapted process model of teaching. Educational assessment, evaluation and accountability*. https://doi.org/10.1007/s11092-021-09373-9
Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
OECD. (2019). *Education at a glance 2019: OECD indicators*. Paris: OECD Publishing. https://doi.org/10.1787/f8d7880d-en
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. https://doi.org/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. https://doi.org/10.3109/0142159X.2013.829912
Rasbash, J., & Browne, W. J. (2008). Non-hierarchical multilevel models. In J. Leeuw, & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 301–334). Springer. https://doi.org/10.1007/978-0-387-73186-5_8.
Razack, S., Faremo, S., Drolet, F., Snell, L., Wiseman, J., & Pickering, J. (2009). Multiple mini-interviews versus traditional interviews: Stakeholder acceptability comparison. *Medical Education, 43*(10), 993–1000. https://doi.org/10.1111/j.1365-2923.2009.03447.x
Rees, E. L., Hawarden, A. W., Dent, G., Hays, R., Bates, J., & Hassell, A. B. (2016). Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A beme systematic review: BEME guide No. 37. *Medical Teacher, 38*(5), 443–455. https://eprints.keele.ac.uk/1489/3/erees_mt_2016.pdf.
Reiter, H. I., Eva, K. W., Rosenfeld, J., & Norman, G. R. (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
Reiter, H. I., Lockyer, J., Ziola, B., Courneya, C. A., & Eva, K. (2012). Should efforts in favor of medical student diversity be focused during admissions or farther upstream? *Academic Medicine, 87*(4), 443–448. https://doi.org/10.1097/ACM.0b013e318248f7f3
Reiter, H. I., Salvatori, P., Rosenfeld, J., Trinh, K., & Eva, K. W. (2006). The effect of defined violations of test security on admissions outcomes using multiple mini-interviews. *Medical Education, 40*(1), 36–42. https://doi.org/10.1111/j.1365-2929.2005.02348.x
Roberts, C., Rothnie, I., Zoanetti, N., & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education, 44*(7), 690–698. https://doi.org/10.1111/j.1365-2923.2010.03689.x
Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., & Tiller, D. (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
Ross, M., Walker, I., Cooke, L., Raman, M., Ravani, P., Coderre, S., & McLaughlin, K. (2017). Are female applicants rated higher than males on the multiple mini-interview? Findings from the university of calgary. *Academic Medicine, 92*(6), 841–846. https://doi.org/10.1097/ACM.0000000000001466
Sabbe, E., & Aelterman, A. (2007). Gender in teaching: A literature review. *Teachers and Teaching: Theory and Practice, 13*(5), 521–538. https://doi.org/10.1080/13540600701561729
Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education, 6*(2), 159–175. https://doi.org/10.1023/A:1011489618208
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. https://doi.org/10.1007/s10459-013-9463-7
Siu, E., & Reiter, H. I. (2009). Overview: What's worked and what hasn't as a guide towards predictive admissions tool development. *Advances in Health Sciences Education, 14*(5), 759–775. https://doi.org/10.1007/s10459-009-9160-8
Skelton, C. (2009). Failing to get men into primary teaching: A feminist critique. *Journal of Education Policy, 24*(1), 39–54. https://doi.org/10.1080/02680930802412677
Thomson, D., Cummings, E., Ferguson, A. K., Moizumi, E. M., Sher, Y., Wang, X., Broad, K., & Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from one Canadian university. *Canadian Journal of Educational Administration and Policy, 121*, 1–23. https://cjc-rcc.ucalgary.ca/index.php/cjeap/article/view/42818.
Uijtdehaage, S., & Parker, N. (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
Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). North-Holland: Elsevier.
Yoshimura, H., Kitazono, H., Fujitani, S., Machi, J., Saiki, T., Suzuki, Y., & Ponnamperuma, G. (2015). Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: A reliability and acceptability comparison. *BMC Medical Education, 15*(1), 1–9. https://doi.org/10.1186/s12909-015-0361-y