Article:

Renaud, J-S, Bourget, M, St-Onge, C et al. (5 more authors) (Cover date: October 2022) Effect of station format on the psychometric properties of Multiple Mini Interviews. Medical Education, 56 (10). pp. 1042-1050. ISSN 0308-0110

https://doi.org/10.1111/medu.14855

TITLE

Effect of station format on the psychometric properties of Multiple Mini Interviews

AUTHORS

Jean-Sébastien Renaud, PhD, Department of Family and Emergency Medicine, Office of Education and Continuing Professional Development, VITAM Research Center, Laval University

Martine Bourget, MD, Department of Psychiatry and Neurosciences, Laval University

Christina St-Onge, Department of Medicine, University of Sherbrooke

Kevin W. Eva, PhD, Centre for Health Education Scholarship, University of British Columbia

Walter Tavares, PhD, Wilson Center, University of Toronto

Alexis Salvador Loye, PhD, Research professional, Laval University

Jean-Michel Leduc, MD, MMEd. Faculty of medicine, University of Montreal

Matt Homer, PhD, School of Education, University of Leeds

Corresponding author

Jean-Sébastien Renaud

1050, avenue de la Médecine, Université Laval, Québec (Québec) Canada G1V 0A6

Phone: 418-656-2131 ext. 402762; Fax : 418-656-2465

Email: jean-sebastien.renaud@fmed.ulaval.ca

ABSTRACT

CONTEXT: Given the widespread use of MMIs, their impact on the selection of candidates, and the considerable resources invested in preparing and administering them, it is essential to ensure their quality. Given the variety of station formats used and the degree to which that factor resides in the control of training programs, that we know so little about format's effect on MMI quality is a considerable oversight. This study assessed the effect of two popular station formats (interview vs. role-play) on the psychometric properties of MMIs.

METHOD: We analyzed candidate data from the first eight years of the "Integrated French Multiple Mini Interviews (IF-MMI) (2010-2017, n = 11 761), an MMI organized yearly by three francophone universities and administered at four testing sites located in two Canadian provinces. There were 84 role-play and 96 interview stations administered, totaling 180 stations. Mixed design ANOVAs were used to test the effect of station format on candidates' scores and stations' discrimination. Cronbach's alpha coefficients for interview and role-play stations were also compared. Predictive validity of both station formats was estimated with a mixed multiple linear regression model testing the relation between interview and role-play scores with average clerkship performance for those who gained entry to medical school (n = 462).

RESULTS: Role-play stations (M = 20.67, SD = 3.38) had a slightly lower mean score than interview stations (M = 21.36, SD = 3.08), p < .01, Cohen's d = .2. The correlation between role-play and interview stations scores was r = 0.5 (p < 0.01). Discrimination coefficients, Cronbach's alpha, and predictive validity statistics did not vary by station format.

CONCLUSION: Interview and role-play stations have comparable psychometric properties, suggesting format to be interchangeable. Programs should select station format based on match to the personal qualities for which they are trying to select.

INTRODUCTION

Medical and other professional schools face considerable challenge in efforts to select candidates who will best meet the requirements of their future roles. Indeed, while it is true that most candidates admitted to medical school (at least in North America) will obtain their degree,[1] a critical aspect of admissions processes is identifying those who are most likely to reflect the ideals of the profession given the length and cost of training[2] and the influence matriculants will eventually have on the nature and quality of health care provision. To achieve that end, medical schools need good indicators of academic performance (ex.: GPA) and of the non-academic qualities considered essential to the practice of modern medicine, including interpersonal skills and professionalism.[3] Many faculties, therefore, invest considerable human and financial resources in the preparation and administration of Multiple Mini Interviews (MMIs), a method of assessing non-academic qualities in a time constrained manner.[3] Originating in Canada,[3] the evidence accumulated over nearly two decades have led to MMIs being adopted by many health professional training programs around the world, including the United Kingdom, the United States, Australia, Saudi Arabia, and Israel.[4]

Given the widespread use of this method of selection, its impact on applicants, and the considerable resources invested in preparing and administering MMIs, it is essential to ensure their quality.[5] Unfortunately, the variability with which MMIs can be implemented creates the risk that admissions committees assume psychometric properties that their particular MMI does not achieve.[6] Indeed, research on MMIs points to a number of factors that may influence their psychometric properties, including the examiners,[7] both in terms of type (standardized actor, student, clinician)[8-10] and assessment style;[11,12] candidate characteristics, such as personality,[13] gender,[10] and ethnic origin;[14] and station-related factors, such as length[15] and the rating scale used.

For example, Uijtdehaage & al.[16] observed that scores' reliability increased by using a normative scoring rubric rather than a 7-point Likert scale. However, this large body of literature remains largely silent regarding the effect of station format.

This is a considerable oversight given that station format is one factor over which programs have control and different formats can have substantial differences in cost. The most two common formats are interview and role-play stations (i.e., "simulated situations.[17] In interview stations, participants interact with an interviewer who is usually also the rater. The questions can be of various types: behavioral, situational, background, knowledge, opinions and attitudes, goals and aspirations, self-description and self-evaluations.[18] In role-play stations, participants interact with an actor in the context of a simulated scenario and are usually evaluated by at least one additional observer.[19] A pair of studies have suggested that, within the context of interview stations, behavioural interviews and situational judgment questions appear to be equivalent[20,21] Regarding the comparison between role-play and interview stations, O'Neill and colleagues[22] found that they do not result in distinct dimensions in a factor analysis and Knorr and colleagues[23] found that women performed marginally better than men in role-play and interview stations compared to group stations. Nowhere, however, have we been able to find a comparison of the psychometric properties of role play vs interview stations. Such evidence would provide an opportunity to adjust the structure of MMIs to optimize the quality of the assessment.

Indeed, there are a variety of reasons why these particular station formats could differ dramatically in their psychometric properties. In another context, traditional interviews performed with employees applying for promotion, it was reported that impression management techniques[24] were used by interviewees less frequently in role play scenarios relative to interviews. The investigators posited that role-playing is more demanding of cognitive resources than being interviewed, plausibly making it more difficult for participants to use impression management to influence raters.[25] From a psychometric perspective, such findings suggest interview methods introduce more construct-irrelevant variance into candidates' scores than role-play methods. Station format could conceivably affect raters' mental workload similarly given the need to play the dual role of rater and interviewer.[26] In role-play stations, the rater typically has no active role in the simulated scenario and can, therefore, keep attention focused on the rating task, which could result in higher quality observations from raters.[27] Finally, role-play stations might introduce error variability into MMIs given the imperfect standardization between actors.[28] Regardless of training, there will inevitably be variability in an actor's performance across candidates and between actors playing the same role for the same station in a circuit or across circuits.[29] In psychometric terms, this means that while interview stations have at least two potential sources of error variance (i.e., raters and situations), role-play stations have at least three potential sources of error variance (i.e., actors, raters, and situations).[28,29] The latter station format could, therefore, be less reliable, all else being equal.

AIM OF THE STUDY

To empirically test these possibilities, the current study was performed to assess the effect of station format (interview vs. role-play) on the psychometric properties of MMIs. More specifically, the study focused on four properties: station difficulty, discrimination, reliability, and predictive power. Those properties were chosen because the weight of evidence required when documenting the validity of an assessment tool differs depending on the intended use.[30,31] MMI results are used to rank candidates, and their score is expected to be a predictor of future non-academic qualities. Therefore, the extent to which the different MMIs differ in difficulty or to which different stations accurately discriminate among candidates, rank them reliably, and predict non-academic performance over the course of their medical studies (particularly during clerkship rotations[32-34]) are clearly important properties.

METHOD

THE INTEGRATED FRENCH MMI (IF-MMI)

The IF-MMI is a French MMI organized yearly by three francophone universities (Université Laval, Université de Montréal, Université de Sherbrooke) offering an MD program. It is administered at four testing sites (Quebec City, Montreal, Sherbrooke, and Moncton) on the same two days. However, data from Moncton were excluded because that site had a very small number of candidates in each administration (sometimes fewer than 10) and tested on only one of the two testing days.

On average, 1,470 candidates are assessed each year. Both days share the same MMI blueprint but contain different stations. Since 2010, the number of stations has ranged between 10 and 12, each lasting 7 minutes. Participants are rated on each station using a 6-option rating scale that was marked using 5-point increments (i.e., candidates could receive 5, 10, 15, 20, 25, or 30 points for their performance). Before selection, scores are standardized on rater and day of testing to cancel out the effect of test site.

The IF-MMI uses the CanMEDS framework[35] as a blueprint by prioritizing effort to assess qualities like professionalism, empathy, communication, teamwork, leadership, adaptability, self-criticism, the ability to exercise informed judgment, critical thinking, social awareness, and openmindedness. The balance of focus and station format adopted to assess such qualities varies depending on the year.

DATA

Participants

We utilized data from the first eight years of the IF-MMI (2010-2017), which was the most up-to-date administrative database available at the beginning of the study. The number of candidates per testing day varied between 581 and 836 and totaled 11,761. The database used contained the following variables: year, testing site, testing day, gender, stations' score, and total score.

Stations

From 2010 to 2015, there was an equivalent number of role-play and interview stations, six of each format on each MMI administration. In 2016, there were three role-play and seven interview stations. In 2017, there were three role-play and five interview stations (with two additional collaboration stations that were excluded from this study). Therefore, between 2010 and 2017, there were 84 role-play and 96 interview stations administered, for a total of 180 stations.

ANALYSES

IF-MMI's factor structure

To test whether or not it was sensible to group stations based on station format for subsequent analyses, we began by conducting confirmatory factor analyses (CFA) aimed at testing the extent to which the data fit a unidimensional or multidimensional structure. Given that stations are expected to overlap (for example, because communication skill is important to all station formats), we were particularly interested in testing a second-order factor model (in which subskills load on a general ability latent trait) and a bifactor model (which hypothesize a general factor onto which all items load AND a compose series of uncorrelated skill-specific factor groupings.[36] We interpreted fit indices in accordance to commonly used guidelines.[37]

Internal structure analyses

Analyses related to internal structure were performed to test if station difficulty (i.e., mean rating assigned), discrimination (item-total correlation), and reliability (Cronbach's alpha) differed according to station format.

Station difficulty

For each candidate, two mean scores were computed: one for interview stations and the other for role-play stations. A mixed design ANOVA was used to study the fixed effect of station format (discussion vs. role-play) and testing site on candidates' scores. The year and site by year interaction were included in the model as random effects to determine the stability of any differences observed.

Station discrimination

The discrimination index was estimated for all stations using item-total correlations, also known as item-rest correlations[38], where the total is calculated excluding the item of interest[39]. The total score used was the candidates' total score on all MMI stations completed on a testing day. A mixed design ANOVA was also used to study the fixed effect of the type of station (discussion vs. role-play) and the testing site on station discrimination. The year and site by year interaction, considered random effects, were included in the model to determine the stability of any differences observed.

Reliability

The IF-MMI overall reliability ranged between 0.62 and 0.76 for the years included in the study. We calculated Cronbach's alpha separately for both station formats for each testing day. This resulted in 32 coefficients (8 years x 2 days x 2 station formats), 16 for the role-play stations and 16 for the interview stations. There were six role-play and six interview stations for MMIs between 2010 and 2015. For the sake of comparability, we used the Spearman-Brown prediction formula[40,41] to adjust Cronbach's alpha for the years 2016 and 2017, estimating what it would have been if there had been six stations for each station format.

Predictive validity analyses

To estimate the predictive validity of both station formats, we selected a subsample of candidates admitted to Université Laval who 1) had at least started their clerkship at the time of the study; and 2) were assessed using the same form (32 items, α=0.92) at the end of their clinical rotations. This subsample (*n*=462) came from three admission cohorts: 2011 (*n*=217), 2012 (*n*=207), 2013 (*n*=38). Predictive validity was estimated by the association of interview and role-play scores with average clerkship performance. A previous study showed that the IF-MMI correlated better with clerkship performance than with the annual progress test and integrative courses used at this institution.[32]

Candidates' mean scores on interview and role-play stations were first correlated with their mean score on clerkship rotations. That is, we aggregated data by station format to simulate an MMI that would consist of only role-play or discussion stations. We also estimated what the correlation coefficients would be upon correction for unreliability of measurement[42] and range restriction[43] because only candidates admitted to Université Laval were included. Next, we used a mixed multiple linear regression model, in which year was considered as a random factor, to test the relationship between interview and role-play scores and average clerkship performance. The regression slopes for interview and role-play scores were compared by performing a t-test on the parameter estimates (beta and standard error) for both interview types to discern whether one of those station formats is more predictive of clerkship average performance.

All analyses were performed using SAS software (SAS Inc., NC, release 9.4) with p= 0.05 treated as the level of significance.

ETHICS

This project was submitted to Université Laval's research ethics committee and was exempted from human ethics review (#2017-220).

RESULTS

FACTOR STRUCTURE

Upon performing unidimensional, correlated factors, second-order factor, and bifactor analyses, the best fit model for the IF-MMI was the bifactor model (i.e., it explained the most variance in the data). In other words, performance on interview and role-play stations was not independent, but was still separable in a manner that suggests role-play and interview stations addressed somewhat different latent constructs. Given that these analyses are not central to the research question and the specific content of individual stations cannot be shared, we have provided examples of the CFA results in an online supplement (Appendix 1).

INTERNAL STRUCTURE ANALYSES

Station difficulty

Role-play stations had a mean score (*M*=20.67, *SD*=3.4), 0.69 points lower than interview stations (*M*=21.36, *SD*=3.1), *F* (2,22 000)=484.34 (*p*<0.01), , *d*=0.2, 95% CI [0.17, 0.23] (Table 1). This effect is small, given effect size conventions and considering that the scale was marked at 5-point increments. The correlation between role-play and interview station scores was r=0.49, 95% CI [0.48, 0.51],*p*<0.01.

The ANOVA showed a significant interaction between station format and test site, *F* (2,22 000)=7.28 (*p*<0.01). The source of this interaction was variable effect sizes (*d*=0.27, 95% CI [0.23, 0.31], for Site 1, 0.18, 95% CI [0.14, 0.22], for Site 2, and 0.21, 95% CI [0.16, 0.26], for Site 3) rather than that the dominant result of role-play stations being marked slightly, but statistically, lower than interview stations was inconsistent. This consistency is further illustrated in Table 1, which reveals the same direction of difference for 7 out of 8 years, the anomalous year showing a particularly small difference.

Table 1. Descriptive statistics for candidates' scores on both station formats

Year	**n**	**Station format**							
		Interview				**Role-play**			
		Mean	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
2010	1400	22.07	2.79	10.83	29.17	21.22	3.12	8.33	30.00

2011	1414	21.73	2.83	10.83	29.17	20.41	3.09	9.17	29.17
2012	1423	21.74	2.96	10.00	30.00	21.51	3.01	9.17	29.17
2013	1389	21.73	3.13	6.67	29.17	20.47	3.44	7.50	29.17
2014	1403	21.41	3.22	10.00	29.17	20.30	3.20	9.17	28.33
2015	1500	21.44	3.20	9.17	30.00	20.80	3.15	10.00	28.33
2016	1482	20.21	2.97	5.71	27.86	20.28	3.94	5.00	30.00
2017	1162	20.45	3.05	6.00	28.00	20.30	3.76	8.33	30.00
All	11 173	21.36	3.08	5.71	30.00	20.67	3.38	5.00	30.00

Min. = Minimum; Max. = Maximum; SD = Standard deviation

Station discrimination index

Table 2 shows descriptive statistics for the discrimination index (D) of interview and role-play stations. Both had a similar average discrimination index (0.32 vs. 0.33, respectively).

Results of the ANOVA indicated that station format was not associated with discrimination power. Neither the station format by site interaction, *F*(2,483)=0.45, *p*=0.64, nor the station format main effect, *F*(1,477)=1.44, *p*=0.23, were statistically significant.

Table 2. Descriptive statistics for the discrimination index (D) for interview and role-play stations

Year	**Interview**				**Role-play**			
	Mean	**SD**	**Min.**	**Max.**	**Mean**	**SD**	**Min.**	**Max.**
2010	0.31	0.07	0.16	0.46	0.36	0.06	0.26	0.49

2011	0.28	0.07	0.11	0.44	0.30	0.07	0.16	0.45
2012	0.29	0.08	0.14	0.46	0.32	0.08	0.13	0.43
2013	0.36	0.07	0.18	0.50	0.37	0.08	0.10	0.51
2014	0.34	0.07	0.20	0.50	0.35	0.08	0.14	0.48
2015	0.33	0.06	0.17	0.46	0.31	0.07	0.16	0.41
2016	0.34	0.07	0.14	0.49	0.31	0.08	0.14	0.46
2017	0.29	0.07	0.12	0.38	0.26	0.07	0.09	0.35
All	0.32	0.01	0.11	0.50	0.33	0.08	0.09	0.51

Reliability

Cronbach's alpha was calculated for both station formats for the two administration days of each year (Table 3). Table 3 illustrates that the mean and median differences between role-play and interview stations using this metric were 0.

Table 3. Cronbach's alpha for role-play and interview stations

Year	Day	n	Role-play	Interview	Difference
2010	1	696	0.61	0.57	0.04
2010	2	776	0.62	0.55	0.07
2011	1	781	0.52	0.52	0.00
2011	2	702	0.50	0.50	0.00
2012	1	789	0.60	0.55	0.05
2012	2	704	0.55	0.54	0.02
2013	1	703	0.62	0.62	0.00

2013	2	710	0.60	0.59	0.01
2014	1	782	0.61	0.62	-0.01
2014	2	710	0.55	0.60	-0.05
2015	1	836	0.54	0.60	-0.06
2015	2	748	0.55	0.58	-0.03
2016	1	832	0.46	0.57	-0.10
2016	2	751	0.63	0.57	0.06
2017	1	660	0.52	0.52	0.00
2017	2	581	0.53	0.54	-0.01
Mean			0.56	0.56	0.00
Md			0.55	0.57	0.00
SD			0.05	0.04	0.05

PREDICTIVE VALIDITY ANALYSES

We calculated the correlations between mean score on clerkship rotations and each of score on role-play stations, score on interview stations, and total score on the MMI. Table 4 shows these correlations by admission cohort, while Table 5 shows the correlations for the entire subsample (all 3 cohorts) along with corrections for unreliability and restriction of range.

Table 4. Correlation coefficients between MMI scores and mean score on clerkship rotations by admission cohort

Year	***n***	**Role-play stations**	**Interview stations**	**MMI**
		r	**r**	**r**
		95% CI	**95% CI**	**95% CI**
2011	217	0.13	0.13*	0.16*

		[-0.01, 0.25]	[0.001, 0.26]	[0.03, 0.29]
2012	207	0.22*	0.22*	0.26*
		[0.08, 0.34]	[0.09, 0.35]	[0.13, 0.38]
2013	38	0.17	0.24	0.25
		[-0.17, 0.46]	[-0.09, 0.52]	[-0.09, 0.52]
Mean		0.17	0.20	0.22

**p* <.05

Table 5. Correlation coefficients between MMI scores and mean score on clerkship rotations

	Role-play stations	**Interview stations**	**MMI**
	r	**r**	**r**
	95% CI	**95% CI**	**95% CI**
Pearson's correlation	0.18 [0.09, 0.26]	0.18 [0.09, 0.26]	0.21 [0.12, 0.30]
Correlation corrected for the unreliability of measurement	0.25	0.25	0.26
Correlation corrected for restriction in range	0.23	0.23	0.32

NB. All correlations are statistically significant at *p* <.05.

Results of the mixed linear regression analysis indicated that the random effect of year was not significant (Wald Z=0.25, *p*=0.40) and, therefore, that a simpler multiple linear regression could be used. The later revealed a collective significant effect between the scores on interview and role-play stations and the mean score on clerkship rotations, *F*(2,459)=10.98, *p*<0.001, R^2 = .05 (Table

6). Furthermore, they had the same predictive power, as indicated by the non-significant difference between their regression coefficients.

Table 6. Results of the multiple linear regression analysis on the mean clerkship rotations' score

Effect	Estimate	*SE*	95% CI		β	*t*	*p*
			LL	*UL*			
Role-play score	0.005	0.002	0.001	0.009	0.13	2.69	0.007
Interview score	0.006	0.002	0.002	0.010	0.13	2.68	0.008

Note. $R^2 = .05$ ($p < 0001$)

DISCUSSION

With this study we aimed to compare the psychometric properties of two widely used MMI station formats (interview vs. role-play). Analyses were focused on four properties: station difficulty, station discrimination, score reliability, and predictive power. Drawing on a large sample size and multiple years of data, we found that the scores emanating from both station formats were largely indistinguishable. More specifically, our results show that candidates scored consistently higher on interview stations than on role-play stations, but the difference was slight whereas station discrimination, score reliability, and predictive power did not vary with station format. These findings suggest that it is important for MMI designers to ensure all candidates within a given competition encounter the same number of role-play relative to interview stations (to insure their MMI is of comparable difficulty given that even small differences can make the difference for individuals near the threshold of acceptance). Beyond that, station format should be selected based on what seems like the best fit for the qualities the selection committee is hoping to measure and the resources available because there appears to be little to no psychometric reason to prioritize one station format over the other.

As noted earlier, there are a variety of reasons that these particular station formats could have differed in their psychometric properties: role-playing could be more cognitively demanding than participating in an interview, which might make it more difficult for candidates to use impression management techniques;[25] having to lead an interview could put some stress on raters' cognitive resources, thereby reducing the mental capacity left for attention and information processing;[26] role-play stations might introduce error variability into MMIs given the imperfect standardization between actors.[28] Our study, however, suggests that either these theoretical possibilities did not materialize, did not do so in a significant way, or that their effect was offset by other factors (e.g., if there was cognitive cost associated with leading an interview it may have been offset by the gains in maintaining attention span and memory associated with being more mentally active).[44]

In practical terms, however, adding our results to those of previous studies of MMI station format[20-23] further reinforces the conclusion that there is little to no psychometric value derived from choosing one format relative to the other. Indeed, results from those studies show that varying the type of interview station or question have no or only a minor effect on score reliability and that the station format does not seem to affect construct validity.[22,23] This reinforces the perspective that sampling as much as possible (i.e., maximizing the number of stations included) is the dominant influence on the value inherent in the use of MMIs.

As such, admissions offices should focus their energies primarily on maximizing the number of stations they can feasibly run and deciding what qualities they will strive to measure. Upon doing so, all evidence suggests that the specific format chosen to create a station focus is immaterial even though we recommend continuing to balance station format across MMI circuit. Most qualities admissions offices seek to measure using MMIs can be operationalized with either role-play or interview stations. One should not fool themselves, however, into believing that any one station captures only one quality given the illusory nature of efforts to claim skill sets are independent of one another.[45] Our confirmatory factor analyses of the IF-MMI supports this by virtue of the bifactor model (defined as a general factor onto which all items load AND uncorrelated skill-specific subfactors) fitting the available data better than other models considered. That suggests that what is captured will be slightly different depending on the station format, but when combined with the psychometric similarities observed, the choice between station format is better based on the operationalisation the admissions office prefers and their capacity to design stations that tap into similar qualities in a variety of ways. In other words, MMI designers can rest easy knowing that the station format they choose will probably not influence their MMI's psychometric properties as much as it influences its face validity.

Face-validity is a weak component of a validity argument, but politically it can have a considerable impact on a measurement strategy's acceptability. In this regard, it is important to try to ensure that stations are not designed that are likely to bias against any particular demographic subgroup and, as previously stated, that each candidate encounters the same station format balance to ensure that small differences in their difficulty do not negatively impact individuals.

We would advise against choosing station format based on the argument that interview stations are more cost efficient, requiring fewer resources, yet result in comparable psychometric properties. The fact that scores on role-play and interview stations were observed to have a correlation of .49 indicates that they share 24% of common variance, leaving room to believe that they capture different things while still relating similarly to clerkship performance (again, as supported by our confirmatory factor analyses).. Unfortunately, not having more outcome variables focusing on distinct competency areas is the primary limitation of the dataset we were able to secure. Given that MMIs constitute a method rather than a standardized test,[21] the degree to which our results can be generalized to other MMIs will depend on their similarity with the IF-MMI. Second, because the proportion of interview and role-play stations was unequal for 2016 and 2017, we had to rely on the Spearman-Brown prophecy formulae to estimate Cronbach's alpha. These results are, therefore, approximations rather than exact values. Third, the subsample used to test predictive validity was limited to students from Université Laval, as we did not have access to the data from the other universities.

CONCLUSION

In conclusion, we found that interview and role-play stations have comparable psychometric properties in the IF-MMI. While we would express caution until further research is conducted to explore whether or not different stations better measure different personal qualities (i.e., if station formats are indeed interchangeable), our findings suggest that it is more important to assure all candidates within a given admissions cycle encounter as many stations as possible while maintaining a balance within each format (role play vs interview) to avoid slight differences in difficulty biasing student selection.

AUTHOR CONTRIBUTIONS

JSR conceptualised the study. JSR and MB acquired the data. All authors contributed to the analysis of the data and the interpretation of the results. JSR drafted the article. All authors revised it critically and contributed to the preparation of its final version. All authors approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We want to thank Dr Claude Labrie, from Laval University, and Dr Christian Bourdy, from the University of Montreal, for their support to the project.

CONFLICT OF INTEREST

None.

FUNDING

This study was funded by the FRQSC (*Fonds de recherche du Québec – Société et culture*).

REFERENCES

1. Association of Faculties of Medicine of Canada. *Canadian Medical Education Statistics.* Ontario, Canada2015.
2. Health Canada. *Overview of the cost of training health professionals.* Official Languages Community Development Bureau;2008.
3. Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. *Med Educ.* 2004;38(3):314-326.
4. Knorr M, Hissbach J. Multiple mini-interviews: same concept, different approaches. *Med Educ.* 2014;48(12):1157-1175.
5. Ali S, Sadiq Hashmi MS, Umair M, Beg MA, Huda N. Multiple Mini-Interviews: Current perspectives on utility and limitations. *Adv Med Educ Pract.* 2019;10:1031-1038.
6. Reiter H, Eva K. Vive la Différence: The freedom and inherent responsibilities when designing and implementing Multiple Mini-Interviews. *Acad Med.* 2018;93(7):969-971.
7. Roberts C, Walton M, Rothnie I, et al. Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Med Educ.* 2008;42(4):396-404.
8. Dowell J, Lynch B, Till H, Kumwenda B, Husbands A. The multiple mini-interview in the UK context: 3 years of experience at Dundee. *Med Teach.* 2012;34(4):297-304.
9. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Acad Med.* 2004;79(6):602-609.

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

11. Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ.* 2010;44(7):690-698.

12. Till H, Myford C, Dowell J. Improving student selection using Multiple Mini-Interviews with Multifaceted Rasch modeling. *Acad Med.* 2013;88(2):216-223 210.1097/ACM.1090b1013e31827c31820c31825d.

13. Jerant A, Griffin E, Rainwater J, et al. Does applicant personality influence Multiple Mini-Interview performance and medical school acceptance offers? *Acad Med.* 2012;87(9):1250-1259 1210.1097/ACM.1250b1013e31826102ad.

14. Moreau K, Reiter H, Eva KW. Research basic to medical education: comparison of aboriginal and nonaboriginal applicants for admissions on the multiple mini-interview using aboriginal and nonaboriginal interviewers. *Teach Learn Med.* 2006;18(1):58-61.

15. Dodson M, Crotty B, Prideaux D, Carne R, Ward A, De Leeuw E. The Multiple Mini-Interview: how long is long enough? *Med Educ.* 2009;43(2):168-174.

16. Uijtdehaage S, Doyle LH, Parker N. Enhancing the reliability of the Multiple Mini-Interview for selecting prospective health care leaders. *Acad Med.* 2011;86(8):1032-1039.

17. Cameron AJ, MacKeigan LD. Development and pilot testing of a multiple mini-interview for admission to a pharmacy degree program. *Am J Pharm Educ.* 2012;76(1):10.

18. Knorr M, Hissbach J, Hampe W. Interviews, Multiple Mini-Interviews, and selection centers. In: Patterson F, Zibarras L, eds. *Selection and Recruitment in the Healthcare*

Professions Research, Theory and Practice. Cham: Springer International Publishing; 2018.

19. Pau A, Jeevaratnam K, Chen YS, Fall AA, Khoo C, Nadarajah VD. The Multiple Mini-Interview (MMI) for student selection in health professions training – A systematic review. *Med Teach.* 2013;35(12):1027-1041.

20. Yoshimura H, Kitazono H, Fujitani S, et al. Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: a reliability and acceptability comparison. *BMC Med Educ.* 2015;15(1):75.

21. Eva KW, Macala C. Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. *Med Educ.* 2014;48(6):604-613.

22. O'Neill LD, Lykkegaard E, Kulasageram K. Intended and unintended test constructs in a Multiple-Mini admission Interview. A validity study. *Dansk Universitetspædagogisk Tidsskrift.* 2019;14(26):66-81.

23. Knorr M, Meyer H, Sehner S, Hampe W, Zimmermann S. 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 Med Educ.* 2019;19(1):243.

24. Yu KYT, Cable DM. *The Oxford handbook of recruitment.* Oxford: Oxford University Press; 2014.

25. McFarland LA, Ryan AM, Kriska SD. Impression management use and effectiveness across assessment methods. *Journal of Management.* 2003;29(5):641-661.

26. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. *Adv Health Sci Educ Theory Pract.* 2013;18(2):291-303.

27. Lievens F, Schollaert E, Keen G. The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *J Appl Psychol.* 2015;100(4):1169-1188.

28. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. Quality assurance methods for performance-based assessments. *Adv Health Sci Educ.* 2003;8(1):27-47.

29. Baig LA, Beran TN, Vallevand A, Baig ZA, Monroy-Cuadros M. Accuracy of portrayal by standardized patients: Results from four OSCE stations conducted for high stakes examinations. *BMC Med Educ.* 2014;14(1):97.

30. AERA, APA, NCME. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association; 2014.

31. Kane MT. Validation. In: Brennan RL, ed. *Educational measurement.* 4 ed. Wesport, CT: American Council on Education and Praeger; 2006:17-64.

32. Renaud J-S, Cantat A, Lakhal S, Bourget M, St-Onge C. Sélection des candidats en médecine : validité prédictive des mini entrevues multiples en contexte francophone. *Pédagogie Médicale.* 2016.

33. Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. *Acad Med.* 2004;79(10):S40-S42.

34. Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple Mini-Interviews predict clerkship and licensing examination performance. *Med Educ.* 2007;41(4):378-384.

35. Frank J, Snell L, Sherbino J. *CanMEDS 2015 physician competency framework.* Ottawa: Royal College of Physicians and Surgeons of Canada;2015.

36. Dunn KJ, McCray G. The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. *Front Psychol.* 2020;11.

37. Thompson B. Confirmatory factor analysis decision sequence. In. *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association; 2004.

38. Lord FM, Novick MR. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Pub. Co.; 1968.

39. Streiner DL, Norman GR, Cairney J. *Health measurement scales : a practical guide to their development and use.* Fifth edition ed. Oxford: Oxford University Press; 2015.

40. Spearman C. Correlation calculated from faulty data. *Br J Psychol.* 1910;3(3):271-295.

41. Brown W. Some experimental results in the correlation of mental abilities. *Br J Psychol.* 1910;3(3):296-322.

42. Spearman C. The proof and measurement of association between two things. *Am J Psychol.* 1904;15(1):72-101.

43. Thorndike RL. *Personnel selection; test and measurement techniques.* New York: Wiley; 1949.

44. Markant DB, Ruggeri A, Gureckis TM, Xu F. Enhanced memory as a common effect of active learning. *Mind Brain Educ.* 2016;10(3):142-152.

45. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med.* 2009;84(3):301-309.