BMJ Open

Feasibility of an automated interview grounded in multiple mini interview (MMI) methodology for selection into the health professions: an international multimethod evaluation

Alison Callwood ,[1] Lee Gillam,[2] Angelos Christidis,[2] Jia Doulton,[1] Jenny Harris,[1] Marianne Piano,[3] Angela Kubacki,[4] Paul A Tiffin,[5] Karen Roberts,[6] Drew Tarmey,[7] Doris Dalton,[8] Virginia L Valentin[8]

To cite: Callwood A, Gillam L, Christidis A, *et al*. Feasibility of an automated interview grounded in multiple mini interview (MMI) methodology for selection into the health professions: an international multimethod evaluation. *BMJ Open* 2022;**12**:e050394. doi:10.1136/bmjopen-2021-050394

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-050394).

Received 18 February 2021
Accepted 20 December 2021

For numbered affiliations see end of article.

Correspondence to
Dr Alison Callwood;
a.callwood@surrey.ac.uk

ABSTRACT

Objectives Global, COVID-driven restrictions around face-to-face interviews for healthcare student selection have forced admission staff to rapidly adopt adapted online systems before supporting evidence is available. We have developed, what we believe is, the first automated interview grounded in multiple mini-interview (MMI) methodology. This study aimed to explore test–retest reliability, acceptability and usability of the system.

Design, setting and participants Multimethod feasibility study in Physician Associate programmes from two UK and one US university during 2019–2020.

Primary, secondary outcomes Feasibility measures (test–retest reliability, acceptability and usability) were assessed using intraclass correlation (ICC), descriptive statistics, thematic and content analysis.

Methods Volunteers took (T1), then repeated (T2), the automated MMI, with a 7-day interval (±2) then completed an evaluation questionnaire. Admission staff participated in focus group discussions.

Results Sixty-two students and seven admission staff participated; 34 students and 4 staff from UK and 28 students and 3 staff from US universities. Good-excellent test–retest reliability was observed at two sites (US and UK2) with T1 and T2 ICC between 0.65 and 0.81 (p<0.001) when assessed by individual total scores (range 80.6–119), station total scores 0.6–0.91, p<0.005 and individual site (≥0.79 p<0.001). Mean test re-test ICC across all three sites was 0.82 p<0.001 (95% CI 0.7 to 0.9). Admission staff reported potential to reduce resource costs and bias through a more objective screening tool for preselection or to replace some MMI stations in a 'hybrid model'. Maintaining human interaction through 'touch points' was considered essential. Users positively evaluated the system, stating it was intuitive with an accessible interface. Concepts chosen for dynamic probing needed to be appropriately tailored.

Conclusion These preliminary findings suggest that the system is reliable, generating consistent scores for candidates and is acceptable to end users provided human touchpoints are maintained. Thus, there is evidence for the potential of such an automated system to augment healthcare student selection.

Strengths and limitations of this study

- ► The underpinning iterative theoretical approach enabled a responsive, dynamic design and development process for a new technology with no known precedent.
- ► The conceptual leap from face-to-face or videoconference facilitated multiple mini-interviews to a fully automated interview and assessment system may present barriers to stakeholders irrespective of the technology and its features.
- ► The multimethod design provided for a diverse set of insights, which have been essential to informing the progression of the technology.
- ► We were unable to assess for potential differential performance within subgroups as this would require a larger sample size.

INTRODUCTION

Global, COVID-driven social distancing restrictions have forced healthcare admissions staff to rapidly adapt to online systems.[1–3] The rate of change has outstripped published evidence, resulting in interview methods with largely unknown efficacy.[4–6] Our responsibilities to ensure inclusive and robust processes have, therefore, never been more challenging to enact.

Prepandemic, candidate selection was predominantly face-to-face using unstructured or structured approaches including panel interviews, group interviews and multiple mini-interviews (MMIs).[7] MMIs are a series of short, focused interactions with a number of different interviewers.[8] This multistation format featuring scenario questions,

tailored scoring proforma and a unidirectional flow of conversation is designed to mitigate against the potential impact of interviewer bias.[8] MMI scenarios focus on random subject areas intended to assess role-defined attributes and values. This makes it more difficult for candidates to anticipate questions and benefit from any prior 'coaching' by preparing answers. MMIs have been shown to be a feasible, acceptable, valid and reliable candidate selection approach across health professions.[9] Nonetheless, MMIs along with other face-to-face methods are understood to be costly, resource intensive and influenced by unintended bias intrinsic to human assessment.[8]

Technology-facilitated interviews aim to alleviate cost and bias issues. However, pre-COVID, there was limited evidence regarding the effectiveness of such approaches to selection into the health professions. While some published example evaluations in this context exist, including the use of Skype-based MMIs,[10] asynchronous MMIs[11] and asynchronous panel interviews,[12] findings were inconclusive and inconsistent. For example, some candidates reported feeling that the ability to fully express themselves was impaired while others considered the absence of an interviewer made the process more objective.[12] The coronavirus pandemic-driven move to online interviews has resulted in single-site evaluations of the use of video-conference technology (eg, Zoom) to facilitate MMIs.[2 6] Findings suggest that online MMIs are feasible and acceptable provided reliable high-speed internet connection is available. Unintended bias remains a concern, with evidence from outside the field of healthcare suggesting that, for example, video backgrounds can influence assessor perspectives.[13]

Beyond health professions, multinational companies describe resource and bias reduction achieved through technology-enhanced interviews using artificial intelligence (AI). This is an advancement from videoconference-facilitated online interviews to incorporate an element of non-human, automated assessment and rating or scoring. Unilever, for example, use AI to analyse candidates' interviews based on facial expressions and word choice. They report a 50% cost reduction and 16% increase in diversity hiring[14] due to improved accessibility to interviews and reduced opportunity for unconscious bias. There is insufficient evidence to draw causal inferences from these observations, but they remain potentially relevant to health professions selection, where facilitating fairness and diversity is an international priority area.[15]

Nonetheless, the use of online technology in healthcare admissions has been exploratory with the acceptability, feasibility and effectiveness yet to be formally established in a range of relevant personnel selection settings. Consequently, the aim of this study was to evaluate the test–retest reliability, acceptability and perceived usability of what we believe is the first known automated interview and assessment system grounded in MMI methodology. Such a study was intended to pave the way for further development and refinement of the system, prior to evaluating, at scale, its performance, in terms of predictive validity and other key properties, such as potential bias.

The system itself is intended to improve cost-efficiency, and reduce unintended, and undesirable, bias associated with human judged assessments. Although the project was initiated prior to the coronavirus pandemic, the potential of a remote, digital interview to overcome the challenge of social distancing restrictions has made it more relevant to personnel selection in recent times. Importantly, the difference between our automated interview and currently adapted online interviews is that it provides for a fully automated interview and assessment system. That is, ratings can be derived using an AI system, as opposed to scores produced by a person using videoconference technology to facilitate a human-assessed interview using MMI scenarios.[2 3]

METHODS

Our automated interview emulates the principles of face-to-face MMIs,[8] where interview content is analysed for the demonstration of role-relevant values and attributes, but not by a human. An advanced, custom-built digital system combining validated 'off-the-shelf' and bespoke technologies uses techniques of natural language processing (NLP) to identify evidence of construct-relevant attributes and values from narrative interview content. A minimum word limit is required to provide an AI with sufficient information to be able to enable this in-depth analysis from automated transcripts of interview responses. Results provided to assessors are intended to help inform selection decisions where the ability to sense check the reasons for allocated scores per attribute/value, per scenario, per candidate makes for a transparent decision-support tool. The system is summarised in figure 1.

Design

The development and evaluation of the automated interview took place in three phases: scoping, pretest and feasibility study between January 2018 and January 2020.

This paper focuses on the outcomes of the feasibility study (April 2019to December 2019) in admissions to Physician Associate (PA) programmes in two UK and one US university. For completeness, prior work is summarised in figure 2.

This dual paradigmatic, dialectical enquiry[16] was underpinned by Olsen and Eoyang' Complex Adaptive Systems (CAS) model.[17] The pragmatic, 'evolving' systems approach enabled refinements to be made from theoretical conception to deployable system where we were open to new insights as they emerged. The iterative nature of this model aligned with meeting the challenges of developing and piloting a new approach for which there was no known precedent.

Participants

UK universities were invited to act as testbed sites from an MMI Expert Group network. In the US, an invitation

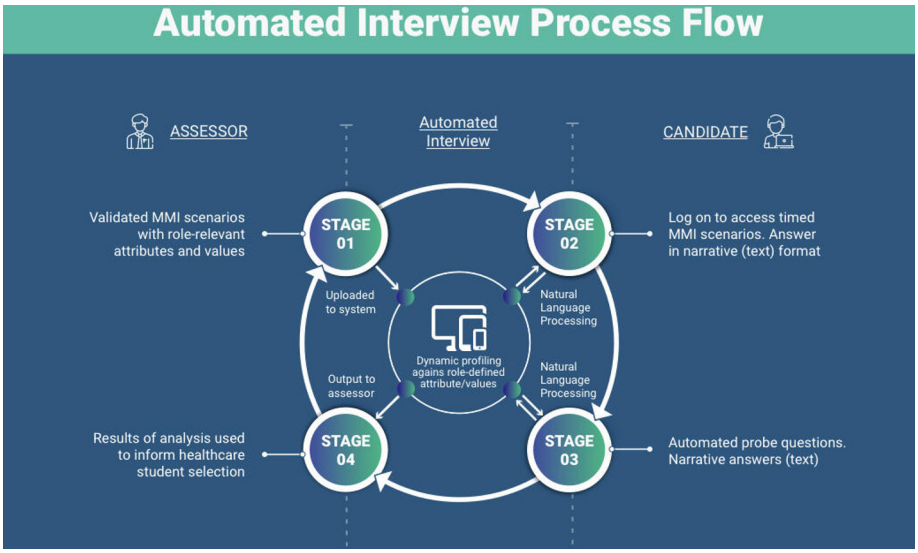

Figure 1 Automated interview process flow.

was sent to PA programme admission leads through a national network.

Admission staff leading PA student selection at collaborating testbed sites worked with the research team to facilitate setting up the automated interview, including supplying site-specific scenarios.

Volunteers to take the automated interview were recruited through non-probability convenience sampling from PA students at collaborating universities between April 2019 and December 2019. In the prepilot, applicants to PA programmes were invited to participate, but this brought challenges to applicants and staff on already stressful interview days. Therefore, study recruitment was broadened to include first-year PA students. This approach aligned with the study aims because, at this stage, we were interested in test–retest performance against successive automated interview scores, deemed an essential step prior to validity testing with 'live' applicants.

Collaborating test-bed sites that were Universities, with first-year PA students, were included. Universities who did not use MMIs and PA students who had been involved in the scoping/prepilot development of the automated interview were excluded from the study.

Patient and public involvement

Past and current health service users were involved in the initial MMI scenario development. This was to ensure that values relevant to patient experience were appropriately accommodated. The results will be disseminated to study contributors.

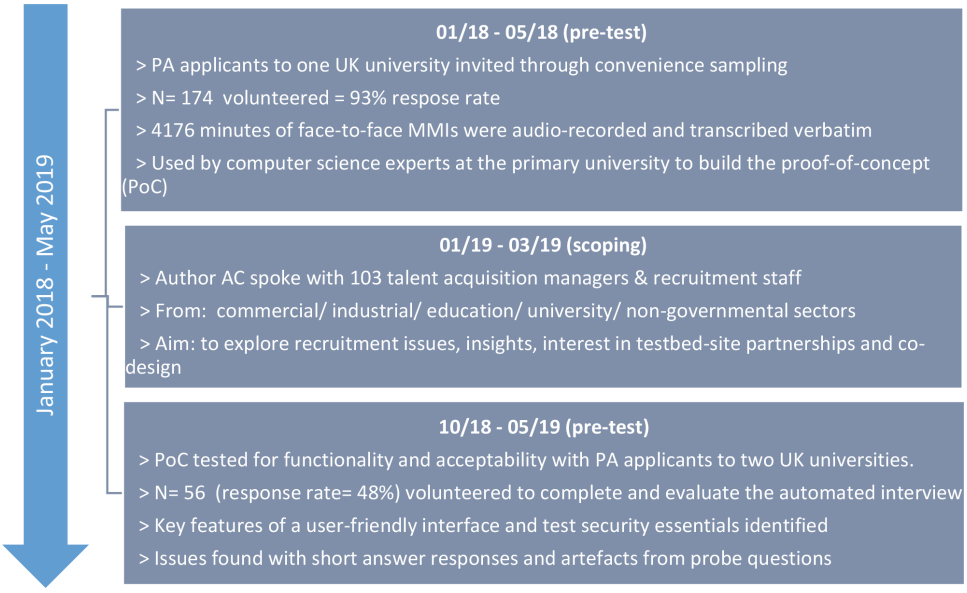

Figure 2 Scoping and pretest activity. PA, Physician Associate.

Data collection

PA students took part at a designated date, time and venue with secure computer access and stable Wi-Fi. They completed four MMI-style scenarios using the automated interview system, writing their answers in text form, allowing a maximum 40 min overall, ±10 min for additional time if required (T1). Using text responses (as opposed to oral) was a pragmatic decision taken at this stage as we were interested in the 'real-time' capability of the system and wanted to be confident in capturing responses. MMIs were site-specific, the content and criteria being replications of the face-to-face scenarios used to interview students during their 'live' selection. Scenario details are withheld for ongoing test security/confidentiality reasons; however, the core attributes/values the MMI scenarios were designed to assess were successfully mapped to the capability of the NLP system to identify coherent topics and themes from the interview transcripts. For test–retest evaluation, volunteers were asked to repeat the same four scenarios 1 week later (±2 days, T2) under similar conditions as T1, thereby minimising carryover effect and the impact of any 'learning'.[18]

Admission staff participated in site-specific focus groups to elicit acceptability perspectives, defined according to Nielsen.[19] An *a priori* topic guide facilitated exploration of their views of the system itself and automation in candidate selection.

To explore usability[19] students completed a study-specific evaluation questionnaire immediately following T2. The questionnaire contained a mixture of closed questions with Likert scales and open text formats as well as demographic data.

Analysis

Automated interview scores for each of the attributes/values were summed at T1 and T2 for each candidate per station and across stations. Descriptive statistics were explored, and test–retest reliability was assessed using the intraclass correlation coefficient (ICC) two-way mixed model.[20] Individual total scores, station total scores, per site and mean scores for T1 and T2 are presented. All analyses were performed in Stata V.16. Scores for attribute/value comparisons at T1 and T2 were also verified using multidimensional distance measures including Manhattan, Euclidean and Cosine measures.[21]

Staff focus group discussions were transcribed verbatim and thematic analysis[22] performed by author MC who was not otherwise connected to the study. This involved reading the transcripts in detail and multiple coding passes using NVivo V.12 (QSR International, USA). Emerging themes were reviewed, and coding conflicts resolved collaboratively with research team member, author AC.

Descriptive statistics of students' characteristics and views are presented. Conventional qualitative content analysis[23] was performed on open-ended questions to elicit students' perspectives of the automated interview software usability.

RESULTS

Sample characteristics

A total of 62 first-year PA students from one US and two UK universities took part (UK1: n=17; UK2: n=17, USA n=28), representing 52% average uptake across sites.

English was the first language of over 70% of student participants. US students differed demographically from those at the two UK sites, with a more even distribution of age groups, a lower proportion identifying as women and being predominantly white. In the UK universities, over 80% of participants were under 30 years of age, over 59% identified as women with greater ethnic diversity. Volunteers had some prior exposure to preselection online assessment systems, including University Clinical

Table 1 User characteristics (n=62 students)

		USA n=28	UK1 n=17	UK2 n=17
English as a first language		22 (78.6%)	12 (70.6%) missing n=1	14 (82.4%) missing n=2
Gender self-identification	Female	12 (43%)	10 (59%)	15 (88%)
	Male	16 (57%)	7 (41%)	2 (12%)
	Prefer not to say/other	0%	0%	0%
Age group	Under 30	50%	82.4%	88.2%
	30 and above	50%	17.6%	11.8%
Ethnicity*	White	53.6%	41.2%	29.4%
	Asian/Asian British	7.14%	41.2%	17.6%
	Black, African, Caribbean, Black British	0%	17.6%	35.3%
	Mixed or multiple Ethnic Groups	35.7%	0%	5.8%
	Other/prefer not to say	3.6%	0%	11.8%

*https://www.ethnicity-facts-figures.service.gov.uk/style-guide/ethnic-groups.

Table 2 ICC between test 1 and test 2 per station, individual and across sites

ICC per station (total scores) at T1 and T2 per site			
Test bed site	**ICC T1 and T2**	**95% CI**	**P value**
US n=26			
Station 1	0.77	0.38 to 0.87	0.001
Station 2	0.6	0.06 to 0.81	0.008
Station 3	0.78	0.49 to 0.89	0
Station 4	0.75	0.45 to 0.89	0
UK2 n=17			
Station 1	0.8	0.30 to 0.87	0.001
Station 2	0.79	0.44 to 0.93	0.001
Station 3	0.91	0.74 to 0.97	0
Station 4	0.74	0.29 to 0.91	0.005
UK1 n=14			
Station 1	0.43	0.79 to 0.82	0.164
Station 2	0.52	0.49 to 0.85	0.098
Station 3	0.73	0.16 to 0.91	0.012
Station 4	0.02	0.16 to 0.73	0.374

ICC of total scores per site					
Test site	**T1 mean (SD)**	**T2 mean (SD)**	**ICC of total scores per student (95% CI)**	**ICC average (95% CI)**	**P value**
US n= 26	104.3 (5.49)	102.9 (7.93)	0.65 (0.35 to 0.82)	0.79 (0.52 to 0.90)	<0.001
UK2 n= 17	100.2 (7.84)	99.7 (6.32)	0.81 (0.55 to 0.93)	0.89 (0.71 to 0.96)	<0.001
UK1 n= 14	97.3 (6.91)	99.2 (9.40)	0.62 (0.15 to 0.86)	0.76 (0.27 to 0.92)	0.007
All sites n= 57	101.4 (7.1)	101.0 (7.9)	0.70 (0.54 to 0.81)	0.82 (0.70 to 0.90)	<0.001

Aptitude Test[24] in the UK and CASPer[25] in the USA, but not a fully automated interview and assessment (table 1).

English was the first language of the seven participating admission staff. Five described themselves as White, one British Asian and one American Asian. There was a gender imbalance with six participants identifying as women and one man, and all were over 40 years of age.

Test–retest evaluation

Complete data including automated interview scores were available for 57/62 (92%) participants (USA n=26, UK1 n=14, UK2 n=17). Two volunteers were unable to finish the retest in USA for personal reasons; attrition at the other sites was due to incomplete/missing data.

Good–excellent reliability was demonstrated at US and UK2 sites with T1 and T2 ICC between 0.65 and 0.81, p<0.001 when assessed by individual total scores (range 80.6–119), station total scores between 0.6 and 0.91, p<0.005, individual site (≥0.79 p<0.001). Mean test–retest ICC across all three sites was 0.82 p<0.001 (95% CI 0.7 to 0.9). Table 2; Manhattan, Euclidean and Cosine measures showed that intracandidate consistency was generally stronger than T1/T2 intercandidate comparisons.

Acceptability

Seven admission staff participated in three focus groups (USA n=3, UK1 n=2, UK2 n=2), representing all those who were approached. The following key themes emerged from analysis of the discussions, illustrated in table 3.

Hybrid or screening tool

Admission staff from all three universities felt that the system could be adopted as an augmentation to in-person interviews in a hybrid approach. There was agreement that selection processes needed some degree of human involvement. It was suggested that this could take several forms including: face-to-face contact with an interviewer for MMI stations; people to supervise and support the automated interview ensuring that technical issues did not disadvantage anxious students; or as an opportunity for prospective students and faculty staff to meet one another in person.

Admission staff also saw a place for the automated interview as a preselection interview tool with potential to mitigate some of the resource costs of conducting face-to-face interviews like MMIs. It was thought that greater focus could then be directed at differentiating between those selected for final interview, reducing the number of interviews having to be conducted by staff.

All admission staff felt that it was important for an automated interview system to be able to measure the same or broadly similar candidate qualities as MMIs in order to be suitable as a direct replacement. This was due to

Table 3 Usability and acceptability evaluation

Student usability questionnaire evaluation		USA (n=28)	UK1 (n=17)	UK2 (n=17)
How helpful did you find the instructions? Median (IQR, skewness)		3.5 (1,–0.651)	3 (1.5,–0.237)	3 (0, 0.051)
How intuitive was the system? Median (IQR, skewness)		4 (1.00,–0.796)	3 (1.00,–0.115)	3 (1.00,–0.855)
How did probe questions relate to overall scenario presented at the beginning? Median (IQR, skewness)		3 (1, 0.584)	2 (1, 1.035)	2 (1.5, 0.054)
Were the probes helpful in allowing you to expand on the answer? Median (IQR, skewness)		2 (0.75, 0.578)	2 (1, 0.741)	2 (0.5, 0.057)
Timer	Less than 3 min	3 (10.7%)	1 (5.9%)	0 (0%)
	3–5 min	13 (46.4%)	9 (52.9%)	3 (17.6%)
	5 min or more	12 (42.9%)	7 (41.2%)	14 (82.4%)
Previous experience of online assessment		13 (46.4%)	2 (11.8%)	3 (17.6%)
Likert scales rated 1–4 with 1 representing a negative statement for example, not helpful at all and 2–4 ranging from least positive, for example, sometimes helpful to most positive, for example, always helpful).				
Student free text comments				
Theme		Illustrative quotes		
Word count		*'The only suggestion I have is to reassess the (minimum) word count' (ID40US)* *'The word count made me less concise' (ID22UK2)*		
Targeted probes		*'Some choices of words to elaborate on did not match the scope' (ID26UK1)* *'It was frustrating when some (probes) were random' (ID31US)*		
Overall		*'I think the idea is great however there is a slight impersonal aspect…' (ID25US)* *'It is very appealing and easy to navigate' (ID31UK2)* *'The programme was smooth' (ID40US)*		
Admissions staff acceptability (focus group discussion)				
Theme	Subtheme detail	Illustrative quotes		
Hybrid or pre-screening tool	Hybrid	*'I can imagine if I said to them, 'here's my plan; next year three of our [MMI] stations are being replaced by this automated thing' [interview] everybody would be like, 'let's do it'.* UK2 *'I would hate to get rid of them(face-to-face MMIs)altogether, but I like this idea of a hybrid'.* USA		
	Pre-selection	*'I think in principal it could be used as a screening tool to try to decrease the number of people we actually interview… It's just become such a massive burden at the moment that anything that would reduce the number of people needed, I think they [admissions staff] would go for it'.* UK2		
	Assessing similar attributes to MMIs	'*But what are the approaches to ensure that what's being measured is the same variables as what's measured in MMI*?' USA		
	Augmentation	*'I'd envisaged it as something you'd use as an MMI station and I would see it very useful as a substitute for some stations, but I think we'd all probably continue to want some face-to-face contact as well'.* UK1		
Objectivity and bias	Inherent bias	'*Whatever their [admissions staff] own inherent biases are, all of that comes into it … so I'm very interested in this idea of bringing in something that doesn't bring in all that bias'.* USA		
	Unconscious bias	'*You can't see if somebody has turned up in jeans, for example, and although we wouldn't discriminate against someone who has turned up in jeans, unconsciously you might'.* UK1		
	Transparency and inbuilt bias	'*I would say if the automated system could prove on some level that it was equal or that it improved equality, because you're taking out the human factor of the interviewer's bias … I think that would be a bonus … because that's one of the issues that everybody complains about is the bias. But I know that anything that is programmed with AI can have the bias of the person who programmed it, so I know that people have concerns about that too.'* UK2		

Continued

Table 3 Continued

Student usability questionnaire evaluation		USA (n=28)	UK1 (n=17)	UK2 (n=17)
Logistics/ technology literacy	Interviewer fatigue	*'So, you've got to do this [MMI station] eight to – [laughing] I don't even know how many times we end up doing it – sixteen times in a day. I know that I'm not the same in my delivery sixteen times in the day. And so, if I have someone who's really struggling, depending on my fatigue level I might score them in a different way'* USA		
	Candidate technology literacy	*'Candidates will manage well, especially with the generation that we are now interviewing, they are very adept at using IT and I don't think it would cause a problem for them as the users, which is very important'*.UK1		
	Staff technology literacy	*'Yes. It's got to be easy to use and understandable…. I think they'd [staff] see it as we are trying to be progressive – I also got a chance to talk to the team across the table as well to find out if it might be a little bit overwhelming … if they said 'I wouldn't know what to expect' or 'I've never been any good with IT' that would be my worry and it's not just older people' [staff]* UK1		
Student perspectives	'Face' of the interview	*'It's important as the 'face' of the interview day … I had a student just tell me that recently; that part of the reason they came here was because our interview process was so much kinder. They felt a warmth here'*.USA		
	Candidate experience	*'Candidate experience is very important, and we are doing what we can'*. UK2		
Cost saving potential	Staff and resource savings	*'We may have three externals on the [interview] day … we actually have a patient rep' as well that we have trained as an assessor. So, he always assesses one of the stations and he's pretty much there every time. And so, we have those outgoings but it's also then the catering, so it's lunch and coffee, tea, blah, blah. So, it's not just the people cost. And then it's the faculty time. So yes, there would definitely be some savings there'*. UK2		

MMI, multiple mini-interview.

an implicit trust in the ability of the face-to-face MMI methodology to enable optimal selection of the desired candidates.

Objectivity and bias

Reducing subjectivity and bias was perceived by admissions staff across sites to be the core benefit and appeal of the system through consistency of evaluation and scoring. As a prescreening tool, it was thought that interviewer burden from volumes of MMIs, understood to exacerbate tiredness and increase the likelihood of bias emerging,[26] could be reduced.

Logistics

Admission staff felt that applicants would manage well with the automated interview interface because of the increasingly widespread use of online selection processes for, for example, part-time jobs. Technology literacy concerns relating to the ability of staff to respond to queries arising with a new system were raised, particularly in the US site.

Student perspectives

Admission staff across universities acknowledged that interviews can be stressful experiences and avoiding technical hiccups was an important priority. A positive applicant experience was thought to be essential. Some concerns were raised, particularly by UK1, that a computer-based interview may not be well received by applicants compared with a face-to-face interview, thereby impacting their attitude towards the university.

Usability

All students who participated in the study (n=62) went on to complete the postautomated interview evaluation questionnaire.

Students were positive overall (median score ≥3 across sites) about the user interface, instructions and the intuitiveness of the system stating it: 'was very appealing', 'easy to use' and 'ran smoothly' (table 3).

When asked about the probe questions in terms of their relevance to the overall scenario and how helpful they were, responses were less positive with a median rating of 2 across all three sites, with the exception of the US university who were more positive. Open-text questionnaire responses suggested that the concepts chosen for students to expand on in the dynamic probing were not always relevant to the scenario or their answer. Only 22.5% of volunteers (n=14) across all sites reported that they felt the probe questions were consistently tailored to their answers.

Volunteers felt that more rather than less time was needed to respond to each of the scenarios, with most indicating a minimum of 3 min was needed. Preferences varied by site with the majority of UK2 university students feeling they needed 5 min or longer.

47%–59% of volunteers, across sites, wrote free-text comments. Over half of these reiterated the need for targeted probe questions or suggestions for a reduced minimum word limit alongside positive feedback as illustrated in table 3.

DISCUSSION

This is the first known evaluation of an online interview and automated assessment grounded in MMI methodology. Previously published examples of online MMIs refer to the facilitated delivery of MMIs using videoconference technologies like Zoom.[2 3 5] However, these are costly, resource intensive and subject to unintended bias intrinsic to human assessments. Our preliminary findings in UK and USA settings provide evidence of good–excellent test–retest reliability as well as acceptability and usability, as long as the system is deployed to augment and not replace human decision-making, and probe questions are appropriately targeted. These insights take on greater significance given the context of the Coronavirus pandemic, resulting in an enforced move to online systems in the absence of robust evidence.

Response rates were above those expected given data collection was prepandemic. This may be illustrative of an emerging acceptance of, or at least familiarity with, technology-augmented interviews already prevalent in recruitment processes outside the field of healthcare.

Our online system provides for a fully automated interview as opposed to a human using videoconference technology to facilitate a human-assessed interview using MMI scenarios. Despite the concept of the automated interview being progressive, admissions staff saw substantial potential to mitigate subjectivity issues associated with human-led interviews through unintended bias and interviewer fatigue.[27] This was based on the consistency of the automated interview in contrast to the perceived nuanced differences between human interviewers conducting MMIs. Generating further evidence to support or refute this is needed. We do not underestimate the potential for inbuilt system biases and recognise essential adherence to best principles in the ethical deployment of trustworthy AI.[28]

Admission staff across test-bed sites were unequivocal that humans should make final decisions about candidate suitability. The automated interview system was considered an augmentation to face-to-face interviews designed for more consistent, less biased evaluation. We acknowledge concerns that a completely automated process with no human-led decision-making may bring unfairness and data protection regulation issues.[29] An automated interview is more remote and abstract from 'real life' and some candidates might find difficulty expressing themselves and communicating effectively without eye contact.[30] Conversely, online interviews have the potential to open up possibilities for applicants removing the need for travel costs and meeting dress code requirements, making selection more accessible and, therefore, fairer. These considerations become more significant in COVID times, where the pandemic is forcing adoption of online methods, sometimes without human touch points or conclusive evidence of fidelity, predictive validity or efficacy. However, since the onset of the pandemic, the population has become more accustomed to online interactions as part of personnel recruitment.[31] Thus, the trend towards remote and digital selection is likely to persist even once coronavirus becomes less of an immediate daily concern. A larger scale study is planned to evaluate potential differences in scoring between current adapted (online) MMIs and our fully automated interview to establish appropriate comparison methods, scoring approaches and predictive validity.

Admission staff were very positive about the possibility of the automated interview to reduce resource costs. There is very limited economic evaluation of online automated interviews in the healthcare selection space that would support these views, and further cost-effectiveness analysis would be beneficial. Outside the field of healthcare, multinationals espouse savings over 80%[14] through online interviews, but we need to be cautious that selection decisions are defensible and do not end up as expensive litigation cases.[32]

How the automated interview was received by applicants mattered to admissions staff, highlighting the need for clear communication to manage expectations and foster optimal applicant performance. Admission staff should consider how they can incorporate human 'touch points' in their online interviews especially as current social distancing restrictions mean personal face-to-face contact is not possible. These can be embedded into the candidate experience by facilitating opportunities to ask questions while online, either during or outside the interview, through virtual campus tours and live webinars/chats.

Student volunteers' overall positivity about the usability of the automated interview is interesting in the context that in the UK, 88% were under the age of 30, and in the USA, almost half had prior experience of online interviews. The iterative codesign, scoping and prepiloting activity appears to have resulted in a system fit for purpose when deployed in an academic setting. The issue of irrelevant probe questions was concerning and reflects the complexity of an automated interaction. It has been addressed in subsequent iterations of our automated interview.

Suggestions to reduce the minimum word limit might impact on the reliability of the linguistic analysis. A new speech capture version of the automated interview now addresses this, as it appears that candidates are more able to articulate their answers when spoken, thereby readily reaching the minimum word limit.

Study limitations

Several limitations should be considered when interpreting our results. The sample size is small[33], particularly UK1 and is limited to one US and two UK Universities, which may not be representative of wider university/student populations. Self-selection bias may have affected outcomes due to the voluntary participation. Over 70% of student volunteers stated that English was their first language. Assessing for potential differential performance in the automated interview for those with English as a first language, compared with those for whom it is not, requires replace 'requires with 'would require' a larger sample size. Awareness is needed when

interpreting the combined ICC as site-specific scenarios and scoring criteria were used. However, these assessed broadly similar constructs and it is reassuring that ICCs by site were similarly high, that is, all ≥0.7.

The acceptability and usability of a new technology may be influenced by other factors we were unable to account for in this study, for example, participants' personal circumstances. This could have influenced their receptiveness to a new innovation, particularly if or how they engaged with and evaluated the system. This will be explored in future studies along with long-term monitoring of potential training bias within the automated system itself.

Study rigour

Author AC conducted the focus groups given her prior experience. A structured interview proforma was used to facilitate discussions to minimise deviation and potential bias. Audio-recordings were transcribed verbatim and 20% was double checked by the research team where 98% accuracy was found. Coding conflicts were resolved with input from AC, and final themes and subthemes were by agreement with all authors.

CONCLUSION

At the time of writing, lack of evidence means that the efficacy of current improvised online interviews is largely unknown. These preliminary findings suggest that our automated interview and assessment system is reliable, generating consistent scores for candidates and is acceptable to end users. There is evidence for the potential of such a system to augment candidate selection, though the perceived importance of maintaining human input was highlighted. These valuable insights are applicable across health professions selection. Further research will focus on evaluating the validity of the automated scores generated against construct-relevant outcomes in future large-scale testing as well as identifying any potential sources of unwanted bias.

Our system significantly advances technology-augmented interviews from videoconference facilitated to a fully automated interview designed to assist admissions staff in making decisions about accepting or rejecting applicants. Conceptually, using technology in this way may be a step too far for some but a welcome innovation for others. Nonetheless, a symbiotic relationship between humans and technology has been forced by social distancing restrictions, and we should be open to understanding possible benefits as well as risks when facing an unknown future beyond COVID-19.

Author affiliations
[1]Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK
[2]Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, UK
[3]Melbourne School of Health Sciences, The University of Melbourne, Melbourne, Victoria, Australia
[4]St George's, University of London, London, UK
[5]Department of Health Sciences, University of York, York, North Yorkshire, UK
[6]Brighton and Sussex Medical School, Brighton, UK
[7]The University of Manchester School of Medical Sciences, Manchester, UK
[8]Department of Family and Preventive Medicine, The University of Utah, Salt Lake City, Utah, USA

Acknowledgements Grateful thanks to the Physician Associate students, collaborating universities and past and current health service users who took part.

Contributors Author AC is guarantor. AC, LG and Ach contributed to the technology development, study design, data collection, analysis, drafting and revision of this paper. JD, DD, DT, KR, AK, VV, PT contributed to the data collection, drafting and revision of this paper. MP and JH contributed to the data analysis, drafting and revision of this paper. All authors reviewed the final manuscript. Ach is no longer affiliated to the University of Surrey but was at the time of design, data collection, analysis and initial drafting of the manuscript.

Funding This work was supported by the United Kingdom Engineering and Physical Sciences Research Council, Impact Acceleration Fund and Innovate UK.

Competing interests Authors AC and LG are co-founders and Ach is an employee of Sammi-Select Ltd, a spinout company from the University of Surrey, UK set up after these data were collected but before this paper was drafted in its final form.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study received a favourable ethical opinion (FEO) from the primary site University Research Ethics Committee (UEC/2017/111/FHMS) and corresponding reciprocal FEO/IRB from collaborating universities in the UK and USA. All participants gave informed consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. Detailed technical information is withheld due to commercial sensitivity.

ORCID iD
Alison Callwood http://orcid.org/0000-0001-9617-909X

REFERENCES

1 Top Universities. University admissions- how will Covid 19 affect my application, 2020. Available: https://www.topuniversities.com/student-info/admissions-advice/university-admissions-covid-19-coronavirus [Accessed 20 Oct 2020].
2 Turpin C, Steele K, Matuk-Villazon O. Rapid Transition to a Virtual Multiple Mini-Interview Admissions Process: A New Medical School's Experience During the COVID-19 Pandemic. *Acad Med* 2021. [Epub ahead of print: 25 May 2021].
3 Ungtrakul T, Lamlertthon W, Boonchoo B, *et al*. Virtual multiple Mini-Interview during the COVID-19 pandemic. *Med Educ* 2020;54:764–5.
4 Sabesan V, Kapur N, Zemanek K, *et al*. Implementation and evaluation of virtual multiple mini interviews as a selection tool for entry into paediatric postgraduate training: a Queensland experience. *Med Teach* 2021:1–8.
5 Cleland J, Chu J, Lim S, *et al*. COVID 19: designing and conducting an online mini-multiple interview (mmi) in a dynamic landscape. *Med Teach* 2020;42:776–80.
6 Kok K, Chen L, Idris F. Conducting multiple mini-interviews in the midst of COVID-19 pandemic. *Korean J Med Educ* 2020;32:281–9.
7 Patterson F, Roberts C, Hanson MD, *et al*. 2018 Ottawa consensus statement: selection and recruitment to the healthcare professions. *Med Teach* 2018;40:1091–101.
8 Eva KW, Rosenfeld J, Reiter HI, *et al*. An admissions OSCE: the multiple mini-interview. *Med Educ* 2004;38:314–26.
9 Yusoff MSB, Yusoff J. Multiple mini interview as an admission tool in higher education: insights from a systematic review. *J Taibah Univ Med Sci* 2019;14:203–40.

10 Tiller D, O'Mara D, Rothnie I, *et al*. Internet-Based multiple mini-interviews for candidate selection for graduate entry programmes. *Med Educ* 2013;47:801–10.
11 Zibarras L, Patterson F, Holmes J, *et al*. An exploration of applicant perceptions of asynchronous video MMIs in medical selection. *MedEdPublish* 2018;7:285.
12 Villwock JA, Bowe SN, Dunleavy D, *et al*. Adding long-term value to the residency selection and assessment process. *Laryngoscope* 2020;130:65–8.
13 An untidy room could cost you your DREAM job, new research finds. Available: https://bmmagazine.co.uk/in-business/an-untidy-room-could-cost-you-your-dream-job-new-research-finds/
14 HireVue. Video interviewing software and platform that makes hiring simple, 2020. Available: https://www.hirevue.com/ [Accessed 10 Oct 2020].
15 United Nations. Sustainable development goals, 2015. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals.html [Accessed 26 Oct 2020].
16 Greene JC, Caracelli VJ, Graham WF. Toward a conceptual framework for mixed-method evaluation designs. *Educ Eval Policy Anal* 1989;11:255–74.
17 Olsen E, Eoyang G. *Facilitating organisational change: lessons from complexity science*. San Francisco: Jossey-Bass, 2001.
18 Salkind N. Test-retest reliability. In: Saljind N, ed. *Encyclopaedia of research design*. Volume 1. London: Sage publications, 2015.
19 Nielsen J. *Usability engineering*. London: AP professional, 1993.
20 Streiner D, Norman G, Cairney J. *Health measurement scales: a practical guide to their development and use*. 5th ed. Oxford: Oxford University, 2014.
21 Szabo F. *The linear algebra survival guide*. Massachusetts, US: Academic Press, 2015.
22 Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;3:77–101.
23 Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;15:1277–88.
24 University clinical aptitude test (UCAT), 2020. Available: https://www.ucat.ac.uk/ [Accessed 12 Jan 2021].
25 Casper, 2020. Available: https://altusassessments.com/casper/how-it-works/ [Accessed 12 Jan 2021].
26 Burgess A, Roberts C, Sureshkumar P, *et al*. Multiple mini interview (mmi) for general practice training selection in Australia: interviewers' motivation. *BMC Med Educ* 2018;18:21.
27 Spuy Vder I, Busch A, Bidonde J. Interviewers' experiences with two multiple mini-interview scoring methods used for admission to a Master of Physical Therapy programme. *Physiother* 2016;68:179–85.
28 European Commission. Ethical guidelines for trustworthy AI, 2019. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [Accessed 26 Oct 2020].
29 Langenfeld T. Internet-based proctored assessment: security and fairness issues. *Educational Measurement: Issues and Practice* 2020;39:24–7.
30 Blacksmith N, Willford J, Behrend T. Technology in the employment interview: a meta-analysis and future research agenda. *Personnel Assessment and Decisions* 2016;2:2.
31 Coronavirus hiring: how recruiters are selecting and interviewing job candidates during the pandemic. Available: https://www.cnbc.com/2020/05/24/how-recruiters-select-and-interview-job-candidates-amid-coronavirus.html
32 Patterson F, Zibarras L. Exploring the construct of perceived job discrimination and a model of applicant propensity for case initiation in selection. *International Journal of Selection & Assessment* 2011;19:251–7.
33 Zaiontz C. 2020 real statistics. Available: https://www.real-statistics.com/reliability/interrater-reliability/intraclass-correlation/icc-for-test-retest-reliability/ [Accessed 12 December 2020].