A randomised controlled trial of the influence of Non-native English accents on Examiners’ scores in OSCEs

CURRENT STATUS: UNDER REVIEW

An Kozato
Barts and The London School of Medicine and Dentistry

ORCiD: https://orcid.org/0000-0002-6955-3653

Nimesh Patel
Barts and The London School of Medicine and Dentistry

Kiyoshi Shikino
Chiba University Hospital

Email: kshikino@gmail.com
Corresponding Author
ORCiD: https://orcid.org/0000-0002-3721-3443

DOI: 10.21203/rs.3.rs-22533/v1

SUBJECT AREAS
   Educational Philosophy and Theory

KEYWORDS
   OSCE, Stereotypes, Bias, Language, Accent, Examiner, Clinical competence
Abstract

Background

Objective Structured Clinical Examinations (OSCEs) are important aspects of assessment in medical education. There are anecdotes that students with non-native English accents (NNEA) may be marked negatively due to unconscious bias. It is imperative to minimise the examiners’ bias so that the difference in the scores reflects students’ clinical competence. Research in shows NNEAs can cause stereotyping, often leading to the speaker being negatively judged. However, no medical education study has looked at the influence of NNEAs in assessment.

Methods

This is a randomised, single-blinded controlled trial. Four videos of one mock OSCE station were produced. A professional actor played a medical student. Two near identical scripts were prepared. Two videos showed the actor speaking with an Indian accent and two videos showed the actor speaking without the accent in either script. Forty-two UK OSCE examiners were recruited and randomly assigned to two groups. They watched two videos online, each with either script, with and without the NNEA. Checklist item scores were analysed with descriptive statistics and simple linear regression model. Global scores were analysed with descriptive statistics and logistic ordinal regression model.

Results

Thirty-two examiners completed the study. The average scores for the checklist items (41.6 points) did not change when the accent variable was changed. Simple linear regression model showed no statistically significant relationship between the accent and the scores (Regression coefficient = 0.032, p = 0.982). For the global scores received by the videos with the NNEA, there were one less ‘Good’ grade and one more ‘Fail’ grade compared to the ones without the NNEA. Logistical ordinal regression model on global scores showed, examiners were more likely to mark the student more negatively (p < 0.0001) but also more positively (p < 0.0001) when the NNEA was present.

Conclusions

Examiners could be biased either positively or negatively towards NNEAs when giving global scores.
Further research is required to consider the nature of this bias. More discussion is warranted to consider how the accent should be considered in current medical education assessment.

Background

Objective Structured Clinical Examinations (OSCEs) are assessment tools commonly used for assessing clinical competence [1, 2]. In an OSCE station, students’ performances are marked by the examiner using a checklist or rating scale [3, 4]. It is stated OSCEs are superior to other assessment methods such as written examination or long cases [3] due to the high construct validity, the standardised cases and marking schemes.

It is important to minimise the examiner bias in OSCEs so that the score differences purely reflect students’ performances. For this reason, formal examiner training is required [5, 6]. Under the Equality Act 2010, all students should be assessed independent of their protected characteristics. Therefore, equality and diversity training is an integral part of the training scheme [5].

Despite the implementation of the examiner training, a systematic review has described that OSCEs were highly variable in their reliability scores [7], especially in the communication skills stations where judgment on the listening skills and cultural competencies tended to be subjective. As the examiner’s measure for the communication is shaped by their cultural background [8], the outcome is likely to be idiosyncratic to the individual.

There have been anecdotes students with non-native English accents (NNEA) were disadvantaged in an OSCE. Although no research has been carried out to establish this issue, studies consistently show students tend to perform worse depending on their protected characteristics such as gender and ethnicity [9-11]. It could be deduced that when examiners make judgements, the decision-making process is influenced by the students’ characteristics and associated social identities.

Studies have consistently shown the correlation between the students’ language status and their performances in OSCE [11-14]. Acculturation was proposed to influence the communication skills in OSCEs. Huhn et al. [15] found that international medical students performed worse especially in the conversational skills stations in the OSCE which was considered as the result of linguistic difficulties.

No investigation was carried out to establish the effect of languages and accents on the examiners.
Outside of medical education, people’s perception towards NNEAs was extensively investigated in speech and psychological studies [16]. An experiment by Fredge [17] in the US showed listeners with little experience in foreign language and phonetics correctly identified the majority of speeches presented to be that of non-native speakers. The variation in consonants, vowels, speeds and speech timings have been identified to play a part in the detection of NNEAs [18].

Studies have demonstrated changes in the attitudes towards speakers with NNEAs compared with native speakers. A speaker with a NNEA was judged more negatively than a native speaker when the message delivered was same [19]. When a NNEA was more prominent, it has been linked to a lower social status [20]. Speakers with NNEAs were marked down in an employment process for high status roles [21]. An experiment by Carlson and McHenry [22] with human resource professionals has shown that when the degree of the perceived ‘accentedness’ of the respective employees were higher, the average employability score decreased.

Overall, there has been a large body of evidence on the effects of NNEAs in assessments. Specifically, when considering medical education there has been a pattern of bias against students or doctors with different nationalities or language status [23]. Medical education research has demonstrated OSCE examiners’ bias against several factors but not on NNEAs. Further investigation is warranted to establish the influence of NNEAs in OSCEs.

Methods

**Study design, Participant recruitment and Ethical approval**

A single-blind randomised controlled experiment was performed to examine a relationship between NNEAs and OSCE scores. The hypothesis was that OSCE examiners scored students with NNEAs lower compared to students with NEAs when the performance is constant.

OSCE examiners in the UK were invited to participate in the experiment via email. Interested individuals contacted the researcher directly to receive the Participants Information Sheet and sign the Consent Form electronically. They needed to have previous formal examiner trainings to be included. Due to the time constraint, sample size of 100 participants at maximum was set as the aimed sample size. They were randomly allocated to group 1 and 2. The alternating group number
was allocated in the order of the consent forms received. The recruitment and randomization process was conducted by the first author.

For the purpose of this study, revealing the true objective of the experiment was considered to predispose the examiners to bias. Therefore, they were simply informed the study evaluated “the assessment reliability in the current OSCE” and that they would be asked to assess two recorded performances of a student in an OSCE station. They were assured that further details would be provided once they completed the study. After completing the task, they were informed the study focused on the effect of NNEAs on OSCE examiners.

Participants were not offered any incentive to complete the study. They were informed they could withdraw from the study at any point. This study was approved by the Queen Mary Ethics of Research Committee (reference code QMERC2018/95).

This study followed the Consolidated Standards of Reporting Trials (CONSORT) reporting guideline (Supplement 1).

**Experiment materials and procedures**

The overall design is summarised in Figure 1. Four videos were prepared as the experiment materials. Each video filmed a mock OSCE station with an identical five minutes scenario. This was a history taking station of leg pain, originally produced for a training purpose. The mark scheme contained checklist items and a global score.

Prior to the recording, two scripts, script 1 and 2, were prepared. The four videos were filmed in the university recording booth. The medical student was played by a professional actor. The actor was female and experienced in performing with and without a NNEA. In this case, an Indian accent was chosen as this corresponded to her ethnicity. The simulated patient was played by a student volunteer. Two videos were filmed for script 1 and 2. The only difference was the presence of the accent in the actor’s speech. The actor was instructed to control everything else such as her body language, facial expression, voice tone and speed.

Two separate websites were created for the two participant groups. They were used as the platforms for viewing the videos and accessing the mark sheet. The websites were created on QMPlus,
Queen Mary’s online learning environment. Each website contained three pages, including introductory information, two videos with the examiner and simulated patient instruction and a weblink to online mark sheets (Figure 2).

The mark sheet was produced online using Wufoo, a digital service for survey production. It contained twenty-one checklist items for which participants choose either good, adequate or inadequate/not done. At the end of the mark sheet, they were asked to give a global rating out of four choices; good, pass, borderline or fail. Another form for the demographic information was produced. Each participant was asked to fill in two identical mark sheets and one demographic information survey. All data was collected online upon submission, after which participants were notified the true objective of the experiment.

The data collection was conducted between 27th March 2019 to 30th April 2019.

**Analysis**

Firstly, we analysed the participants’ demographic information in two groups to see if there was any significant difference between the groups.

The checklist items and the global ratings were analysed separately using descriptive and inferential statistics. Descriptive analysis was carried out using Microsoft Excel and inferential analysis was carried out with SPSS.

The grades for the checklist items were converted into numerical scores for the purpose of the analysis (Table 1). Since there were twenty-one items, all of the numerically converted grades in each marking were summed up. These checklist items score sums, referred as ‘checklist sum’ below, were treated as one variable in the analysis.

The checklist sums given to videos with and without the NNEA were first analysed by descriptive statistics as follows;

The mean values were calculated and compared
The difference between the mean values were analysed with independent samples t-test
The dispersion of the checklist sums were visualised using box plots and compared using standard deviation
The individual examiner’s change in the checklist sum from the first to the second video was visualised by a line chart for each group. This was carried out to see whether there was a pattern in how the examiners scored the first and second videos.
To identify a relationship between the NNEA and checklist sums, inferential statistics were used. We chose simple linear regression model.

The global scores were given in a categorical form in which examiners chose either ‘Good’, ‘Pass’, ‘Borderline’ or ‘Fail’. They were first analysed by descriptive statistics as follows;

Number ‘Good’, ‘Pass’, ‘Borderline’ or ‘Fail’ given to the videos with and without NNEA was counted. This was visualised using a bar chart.

The individual examiner’s change in the global score from the first to the second video was visualised by a line chart. This was carried out to see whether there was a pattern in how the examiners scored the first and second videos.

Inferential statistics was then used to identify a relationship between the NNEA and the global score.

Ordinal logistic regression modeling was used.

Results

Participants

OSCE examiners in the UK were recruited between 27th March 2019 to 10th April 2019. Forty-two examiners were recruited and randomly assigned to group 1 or 2. Thirty-two examiners completed the study. There were fifteen and seventeen examiners in group 1 and 2 respectively. Nine examiners did not complete the study but the reasons were not given. One examiner was excluded from the analysis as the inclusion criteria were not met. None withdrew from the study after the data submission.

Three examiners marked the first video they watched twice in error. For these responses, the first responses were used for the analysis since the second response could be biased due to the previous knowledge of the video contents.

The demographics of the examiners are summarised in Table 2. The overall makeup of the examiners in each group were similar. Although most examiners received equality and diversity training, about half of the examiners in each group received unconscious bias trainings.

Checklist items

Descriptive statistical analysis

The mean value of the checklist sums for videos with and without the NNEA were 41.6 ± SD points v 41.6 ± SD, respectively (Table 3). Independent samples t-test for equality of means showed there was
no statistically significant difference in the mean scores ($p = 0.982$). The dispersion of the scores was visualised using box plots (Figure 3).

The interquartile ranges for both video types were similar. However, the checklist sums for videos without the NNEA was more positively skewed than videos with the NNEA.

The change in the scores given to the first video to the second video was visualised below for group 1 (Figure 4) and group 2 (Figure 5). There were variations in the pattern in which the checklist sums changed for group 1 (Table 4) and group 2 (Table 5). More than 50% of the examiners in group 1 scored the first video A higher than the second video B. Only 33% of the examiners in group 2 scored the first video C higher than the second video D.

**Inferential statistical analysis**

As set out in the above sections, simple linear regression model was used to analyse the relationship between the accent and checklist sums. The coefficient $b$ and the $p$ value were calculated (Table 6). Since the $p$ value is higher than 0.05, there is no statistically significant relationship between the two variables.

**Global scores**

**Descriptive statistical analysis**

The number of the global scores given to videos with and without the NNEA were counted and visualised (Figure 6). Equal number of ‘Pass’ and ‘Borderline’ grades were given. One and zero ‘Good’ grade was given to videos without the NNEA and videos with the NNEA respectively. Four and five ‘Fail’ grade was given to videos without the NNEA and videos with the NNEA respectively.

The individual examiner’s change in the global scores given to the first videos A and C (NNEA present) and second videos B and D (NNEA absent) was visualised (Figure 7). Twenty examiners gave same scores. Seven examiners increased their scores and four examiners decreased their scores.

**Inferential statistical analysis**

Ordinal logistic regression analysis was performed to identify the relationship between the NNEA and the global scores (Table 7). ‘Pass’, ‘Borderline’ and ‘Fail’ were analysed by comparisons against ‘Good’ as the baseline.
For the global score ‘Fail’ the exponential of the coefficient was 0.15. This could be interpreted as the videos with NNEA were 85% less likely to receive a higher score. This finding was statistically significant (p < 0.0001).

For ‘Borderline’, the finding was not statistically significant (p = 0.760), indicating there was no influence of the NNEA on the global score.

For ‘Pass’ the exponential of the coefficient estimate was 55.09. This means the odds for a video with NNEA receiving higher score was 55.09 times than the odds for a video without NNEA. This finding was statistically significant (p < 0.0001).

Since there was only 1 marking with score ‘Good’ it was considered to have affected the validity of the analysis. Second ordinal regression analysis was therefore conducted (Table 8). The score of ‘Good’ was removed from the data set and the ‘Pass’ scores were used as the baseline for the comparison.

For ‘Fail’ the exponential of the coefficient was 0.16. This could be interpreted that a video with NNEA was 84% less likely to receive a higher score. This finding was statistically significant (p < 0.0001).

For the global score ‘Borderline’, the finding was not statistically significant (p = 0.960), indicating there was no influence of the NNEA on the global score. The analysis could not be conducted for ‘Pass’ since this was used as the baseline. This analysis showed same pattern for ‘Fail’ and ‘Borderline’, confirming the consistency in the relationships.

Discussion

**Effects of the accent on the grades**

This study demonstrates that the NNEA could either positively or negatively affect the global scores in OSCEs at a statistically significant level. However, there was no evidence that the NNEA influenced the scores for checklist items. This indicates that the accent could affect the global score and hence the standard setting process but not necessarily impact the assessment of the individual student.

All analysis indicated that the checklist sums were not affected by the accent variable. They are easier to observe than domain based items and the decisions to award a mark is not influenced by the examiner’s subjective judgment to the same degree as the domain based items or global scales.
This could have led to the consistency in the checklist sums when the accent variable was changed. When examiners made judgments on the global scores i.e. the total impression of the student, they were required to decide whether the student was competent overall. This process would be guided by the idiosyncratic experience formed from the cultural and professional norm. This could lead to the stereotype activation while a general impression was formed during the marking. The effect of accents on the judgment has been supported by previous psychological research [24] but never been identified in medical education.

This study presented interesting data regarding the direction of the disparity in the global scores. It could be speculated that the NNEA activated the examiners’ negative stereotypes as described in the literature [19]. However, this is contradicted by the data suggesting the NNEA also increased chance of receiving positive global scores. One explanation for this is that although the accent could trigger a bias, the mechanism behind was not exclusively negative out-group categorisation.

People are very sensitive to speech differences and can distinguish someone to have a ‘foreign’ accent quickly [17]. On top of the initial assignment of the stereotypical membership status, it is necessary to consider what happens beyond. It has been demonstrated that when native listeners encountered speakers with NNEAs, they accommodated to the features of the accented speeches [25]. Native listeners expected non-native speakers to be slower than native speakers. They preferred speeches with the expected rate to speeches with a much faster rate.

Suppose that examiners in this study recognised the NNEA, they might have accommodated to the difference in the accent while they were marking. This accommodation could assist examiners to listen to the student well even though the intelligibility was reduced, reducing the effect of the NNEA. Another study found that native listeners were likely to conclude that any divergence of speech patterns is because speakers were non-native [26]. This implies that even when a non-native person makes a communication error, it is more likely to be associated with their language background rather than their ability. In this study, examiners could have perceived an unsatisfactory communication as consequences of the student being non-native. If this was the case, then the NNEA could have created a positive bias.
This is a slightly different mechanism to the stereotyping activation that has been discussed so far. It would be of interest to see how much the active effort to understand a speech with NNEA and a prior recognition of the non-native status interacts with the stereotyping.

Another possibility could be that the activated stereotype could have been associated with positive characteristics such as empathy or friendliness [24]. This is an example of positive associations to the stereotypes.

Alternative explanation is the contradictory effects of consciously held stereotypes. If an examiner initially perceives a student with a NNEA as ‘foreign’ and therefore ‘underperforming’, what would happen if the student’s performance disconfirms this stereotype? There has been relevant research in the field of the consumer business. When a product performance is higher than the pre-purchase expectation, it leads to a positive consumer response [27]. If this theory is applied to this study, it could be considered students with NNEAs could be assessed more positively than native peers when both perform similarly well.

**Implications**

Although this study was a small-scale pilot experiment, the findings showed a disparity in the global scores due to a NNEA. This could lead to further discussion not only on the matter of OSCE reliability and bias but also carry a wider implication for the increasingly diverse medical student population in the UK.

The experiment showed the accent could lead to both negative and positive bias. As previously mentioned, it is not straightforward to decide whether marking down due to NNEAs could be justifiable or not. It could be argued that decreased comprehensibility due to NNEAs [28] impairs the communication with a patient. However, a line between the accent being a real communication barrier or a ‘perceived’ issue for the examiner is hard to draw.

Meanwhile, the presence of a positive bias presents an interesting implication. No previous research described medical students with NNEAs could be at an advantage. This trend needs to be addressed in a similar manner to the negative bias. In other words, did the examiner feel having an accent assisted the clinical examination/interviews or did they unconsciously favour the accent?
This study also initiates further discussion on what the consensus is among the medical education community regarding the communication variability within a standardised clinical assessment.

Whether the observed positive and negative bias was unconscious or not, it is crucial to consider how the examiners and educational authorities should treat this disparity.

This study also has implication on current and prospective medical students. Identification of the accent bias leads to non-native students to perceive that they could have either an advantages or disadvantages despite being similarly competent to their native peers. Burgess et al. [29] described a ‘stereotype threat’. When a student recognises the existence of a negative stereotype towards his/her characteristics, this leads to an unconscious hindrance in the performance. Unlike the negative bias towards NNEAs, positive bias was not previously acknowledged. It is difficult to expect what implication this finding would have on students. It might be a case where some with socially favoured accents would maintain their communication and speech styles whereas others with accents with negatively associated factors would suppress their original speech style in OSCEs.

It is not just the medical students’ diversity that needs further consideration. Of all doctors who obtained full registration with the General Medical Council (GMC), 58% qualified outside of the UK [30]. The proportion of international medical graduates practicing in the UK has increased since the 1960s [30].

These statistics mean that there could also be an increase in the international medical graduate examiners. This could affect how NNEAs influence the examiners considering the stereotyping is dependent on the individual’s social identity [31]. How the communication divergence is evaluated is also different depending on the native/non-native status of a listener [26]. It might be a case that the effect of NNEAs in the current OSCE would change with the increase in abroad-trained examiners.

Similarly, there has been a rise in the migrant population in the UK [32]. The significance of NNEAs in patient communication may change, ultimately leading to a shift in the definition of clinical competence.

**Limitations**

Due to a time restriction, the number of the examiners in the study was low. The analysis
demonstrated statistically significant findings in the global scores. But the number of ‘Pass’ and ‘Borderline’, two most dominant grades given, was equal for both videos with and without the NNEA. Therefore, it is likely that the statistical findings have risen from the difference of one count in the ‘Good’ and ‘Fail’ grades. The possibility that this was due to chance would be difficult to exclude even with the use of the p value, considering the small sample size.

The global score is not the direct determinant of a student’s outcome in OSCE [33] but is used as the markers in the borderline regression method to set a pass mark. Borderline regression is not reliable when the data set is drawn from a small sample of less than 50 cohorts [34]. Due to the low number of examiners involved in this study, it is difficult to say the observed change in the global scores due to accent would have any actual effect on the OSCE outcome. Conducting a similar experiment with more examiners would be important in identifying a reliable relationship between NNEAs and OSCE marks.

Although the examiners were blinded from the study aim, some might have deduced it during the experiment. There were only two videos for each examiner. It would be possible for them to notice the speeches in the videos were different. This could have produced bias as more conscious effort would be taken to minimise the marking variation.

As shown in the result, the pattern of the score change for checklist sums was different for group 1 and group 2. It could be argued that the order the examiners watched scripts 1 and 2 influenced how the performance was perceived. This might have been a co-founding variable.

Using more videos with variations in accents, student characteristics and scripts would improve the blinding. The effect of script order would also be reduced. This was not feasible in this study due to the resource constraint. Examiners could further be asked after the submission of the marks to comment on what they thought the study was investigating. This information would aid in evaluating the validity of the study.

In this study, the language background of the examiners were not asked. This information would be valuable because the native/non-native status is highly relevant to how people assess NNEAs. It was not possible to look at the effect of the demographics information in this study since two
markings of from one examiners were used in the analysis. It would be possible to analyse the demographics impact by grouping the data based on smaller subsections of the participant group. It would allow demographics information and accent variables to be treated as isolated variables. More examiners are required to conduct this analysis reliably.

It would also be valuable to investigate the effect on the simulated patient evaluation. The mark scheme used in the experiment had one item for empathy which the simulated patient was asked to contribute to. Analysis of this item would explore the influence of NNEAs on the assessment marks comprehensively. It would also show the influence of accents on the wider public.

In this study, one type of NNEA was compared to NEA. Researchers suggested that perceived stereotypes could be markedly different depending on the type of the accent [22]. Therefore, further research into the effect of several types of NNEA on examiners is required.

Quantitative study alone would not provide an insight into why the NNEAs might be influencing the examiners. Study methods such as interviews should be considered to explore this issue.

Another point to consider is how NNEAs could be interpreted in light of clinical competence. It would benefit students, medical educators and patients to discuss how NNEAs in the clinical context are viewed by all stakeholders. This would inform the training process of OSCE examiners the future.

Conclusion

This study provides an insight into the influence of the NNEAs on OSCE examiners. Although there had been anecdotes regarding the bias against non-native students, no research into this matter was present in medical education. The findings of this study would provide a starting point for further investigation.

The findings in this study have shown that students could be more positively or negatively evaluated when they have NNEAs. This is of interest considering that previous studies often showed the presence of negative influences caused by accents. There are several possible explanations why this might occur, but further exploration is necessary to identify the underlying mechanisms.

It is important to acknowledge that the finding only applies to the global score which is concerned with the standard setting, rather than the individual student. Therefore, it would not be feasible to
conclude that students with NNEAs could be directly influenced by the observed trend. Further study is required to clarify whether the observed effect of the accent have truly significant impact on the OSCE reliability.

This study highlights an important topic for current medical education in the UK, given the increasingly diverse population. Clinical competence is a complex and dynamic concept. It could change based on the environment medical professionals work in and the people they interact with. Discussions on the role of accents in not just OSCEs, but in wider field of medical education would lead to the provision of better student education for the future.

**Abbreviations**

GMC: General Medical Council

NEA: Native English Accent

NNEA: Non-native English Accent

OSCE: Objective Structured Clinical Examination

**Declarations**

Ethics approval and consent to participate

This research was reviewed and fully approved by the Queen Mary Ethics of Research Committee in February 2019 with a reference code of QMERC2018/95. The participant information sheet and the consent form were read by the participants before they signed the consent form.

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

**Funding**

Funding was granted from Barts and the London School of Medicine and Dentistry Queen Mary
University of London, for the purpose of creating the experimental materials.

Authors' contributions

AK and NP contributed in the conception of the hypothesis and designed the study. AK produce the experimental material, collected the data and performed a statistical analysis of the data. NP reviewed the execution of the study experiment. AK and NP interpreted the data. AK was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgement

I like to thank Mr Mansor Reizaian for assisting with the data analysis and Ms Karin Fernandes for assisting with the experiment materials. I would like to express my gratitude for Ms Radhika Aggarwal for her assistance in the experiment videos. I would like to thank Mr Sajeel Din for his support with the experiment.

References

1. Norman G. Research in medical education: three decades of progress. BMJ. 2002;324:1560-2.

2. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1996;1:41-67.

3. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Med J. 2011;80:127-32.

4. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. BMJ. 1975;1:447-51.

5. General Medical Council. Assessment in undergraduate medical education: advice supplementary to Tomorrow’s Doctors. London: General Medical Council; 2009. https://www.gmc-uk.org//media/documents/Assessment_in_undergraduate_medical_education___guidance_0815.pdf_56439668.pdf. Accessed 29 April 2019.

6. Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical
Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. Med Teach. 2013;35:1447-63,

7. Brannick MT, Erol!Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011;45: 1181-9.

8. Schwartzman E, Hsu DI, Law AV, Chung EP. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. Patient Educ Couns. 2011;83:472-7.

9. Haq I, Higham J, Morris R, Dacre J. Effect of ethnicity and gender on performance in undergraduate medical examinations. Med Educ. 2005;39:1126-8.

10. McManus IC, Richards P, Winder BC, Sproston KA. Final examination performance of medical students from ethnic minorities. Med Educ. 1996;30:195-200.

11. Wass V, Roberts C, Hoogenboom R, Jones R, Van der Vleuten C. Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. BMJ. 2003;326:800-3.

12. Schoonheim!Klein M, Hoogstraten J, Habets LLMH, Aartman I, Van der Vleuten C, Manogue M, Van der Velden, U. Language background and OSCE performance: a study of potential bias. Eur J Dent Educ. 2007;11:222-9.

13. Liddell MJ, Koritsas S. Effect of medical students’ ethnicity on their attitudes towards consultation skills and final year examination performance. Med Educ. 2004;38:187-98.

14. Mann C, Canny B, Lindley J, Rajan R. The influence of language family on academic performance in year 1 and 2 MBBS students. Med Educ. 2010;44:786-94.

15. Huhn D, Lauter J, Ely DR, Koch E, Möltner A, Herzog W, Resch F, Herpertz SC, Nikendel C. Performance of international medical students in psychosocial medicine. BMC Med Educ. 2017;17:111.
16. Munro MJ. A primer on accent discrimination in the Canadian context. TESL Canada Journal. 2003;20:38-51.

17. Flege JE. The detection of French accent by American listeners. J Acoust Soc Am. 1984;76:692-707.

18. De Jong KJ. Sensitivity to Foreign Accent. Acoust Today. 2018;14:9-16.

19. Raisler I. Differential response to the same message delivered by native and foreign speakers. Foreign Lang Ann. 1976;9:256-9.

20. Brennan EM, Brennan JS. Accent scaling and language attitudes: Reactions to Mexican American English speech. Lang Speech. 1981;24:207-21.

21. Kalin R, Rayko D. Discrimination in evaluative judgments against foreign- accented job candidates. Psychol Rep. 1978;43:1203-9.

22. Carlson HK, McHenry MA. Effect of accent and dialect on employability. J Employ Couns. 2006;43:70-83.

23. Louis WR. Lalonde RN. Esses VM. Bias against foreign born or foreign trained doctors: experimental evidence. Med Educ. 2010;44:1241-7.

24. Cargile AC. Speaker evaluation measures of language attitudes: Evidence of information-processing effects. Language Awareness. 2002;11:178-91.

25. Munro MJ, Derwing TM. Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. Studies in Second Language Acquisition. 2001;23:451-68.

26. Alexander ST. Sincerity, Intonation, and Apologies: A Case Study of ai EFL and ESL Learners. [Dissertation], Bloomington, IN: Indiana University. 2001.

27. Westbrook RA, Reilly MD. Value-Percept Disparity: an Alternative to the Disconfirmation of Expectations Theory of Consumer Satisfaction. In Bagozzi RP, Tybout AM, Abor A, editors. NA - Advances in Consumer Research Volume 10. MI:
Association for Consumer Research, 1983. p.256-61.

28. Lev-Ari S, Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility. J Exp Soc Psychol. 2010;46:1093-6.

29. Burgess DJ, Warren J, Phelan S, Dovidio J, Van Ryn M. Stereotype threat and health disparities: what medical educators and future physicians need to know. J Gen Intern Med. 2010;25:169-77.

30. Goldacre MJ, Davidson JM, Lambert TW. Country of training and ethnic origin of UK doctors: database and survey studies. BMJ. 2004;329:597.

31. Allport GW, Clark K, Pettigrew T. The nature of prejudice. Reading: Addison- Wesley Pub. Co; 1954.

32. Manacorda M, Manning A, Wadsworth J. The impact of immigration on the structure of wages: theory and evidence from Britain. J Eur Econ Assoc. 2012;10:120-51.

33. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000;75:267-71.

34. Homer M, Pell G, Fuller R, Patterson J. Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. Medical Teach. 2016;38:181-8.

Tables

**Table 1:** Categorical grades for checklist items converted into numerical scores

| Checklist item | Categorical grades | Numerical scores |
|----------------|--------------------|------------------|
| Good           | 3                  |
| Adequate       | 2                  |
| Inadequate/not done | 1               |

Legend: The ordinal grades of good, adequate and inadequate/not done for the checklist items were converted into numerical values 3, 2 and 1 respectively.

**Table 2:** Participant demographics
Characteristic

| Received unconscious bias training? | Group 1 | Group 2 |
|------------------------------------|---------|---------|
| Yes                                | 9 (60)  | 9 (52)  |
| No                                 | 6 (40)  | 8 (48)  |

| Received equality and diversity training? | Group 1 | Group 2 |
|-------------------------------------------|---------|---------|
| Yes                                       | 14 (93) | 14 (82) |
| No                                        | 1 (7)   | 3 (18)  |

| Gender |
|--------|
| Male   | 12 (80) |
| Female | 3 (20)  |

| Ethnicity |
|-----------|
| White     | 11 (73) |
| Asian     | 3 (20)  |
| Black     | 0 (0)   |
| Other     | 1 (7)   |

| Level of training |
|-------------------|
| FY1               | 0 (0)   |
| FY2               | 0 (0)   |
| Core Training     | 0 (0)   |
| Specialty training| 1 (7)   |
| Consultant        | 11 (73) |
| Other             | 3 (20)  |

| Years of experience as an OSCE examiner |
|----------------------------------------|
| < 5                                    | 7 (47) |
| 5 - 10                                 | 8 (47) |
| 11 - 15                                | 8 (29) |
| 16 - 20                                | 1 (6)  |

Legend: The demographic information of the participants in each group was summarised.

Table 3: Mean and standard deviation values for checklist sums

| Checklist sum | With NNEA | Without NNEA |
|---------------|-----------|--------------|
| Mean          | 41.6      | 41.6         |
| Standard deviation | 5.3     | 5.8          |

Legend: The mean values for the checklist sum were 41.6 for performances with and without the NNEA. The standard deviations for performances with and without the NNEA were 5.3 and 5.8 respectively.

Table 4: Pattern of score changes from the first video to second video in group 1

| Change in score from the first video (Video A) to the second video (Video B) |
|-------------------|-------------------|-------------------|
| Increase          | Decrease          | Unc               |
| Number of examiners| 7                 | 8                 | 1                 |
| Range of the score change | 1 - 12            | 1 - 7             | 0                 |

Legend: 7 examiners rated the second video higher in the range of 1 to 12 points. 8 examiners rated the second video lower in the range of 1 to 7 points. 1 examiner gave same score to both videos.

Table 5: Pattern of score changes from the first video to second video in group 2

| Change in score from the first video (Video C) to the second video (Video D) |
|-------------------|-------------------|-------------------|
| Increase          | Decrease          | Unc               |
| Number of examiners| 5                 | 10                | 0                 |
| Range of the score change | 2 - 12            | 1 - 9             | -                 |

Legend: 5 examiners rated the second video higher in the range of 2 to 12 points. 10 examiners rated the second video lower in the range of 1 to 9 points. 0 examiners gave same score to both videos.
rated the second video lower in the range of 1 to 9 points.

**Table 6:** Statistical analysis results for checklist sums

| Calculated values | Regression coefficient | P value |
|-------------------|------------------------|---------|
|                   | 0.032                  | 0.982   |

Legend: The regression coefficient for the relationship between the checklist sum and the NNEA was 0.32 at statistically insignificant level (p = 0.982).

**Table 7:** Data from ordinal logistics regression analysis

| Coefficient estimates | Exponential values of the coefficient estimates | P value | Confidence interval |
|-----------------------|-----------------------------------------------|---------|---------------------|
| Fail                  | -1.883                                        | 0.15    | <0.0001             |
| Borderline            | -0.107                                        | 0.89    | 0.760               |
| Pass                  | 4.009                                         | 55.09   | <0.0001             | -2.748 – -1.018 |

Legend: Coefficient estimates for ‘Fail’ score was -1.883 at statistically significant level (p < 0.0001).

For ‘Borderline’, the coefficient was -0.107 at statistically insignificant level (p = 0.760). For ‘Pass’, the coefficient was 4.009 at statistically significant level (p < 0.0001).

**Table 8:** Data from secondary ordinal logistics regression analysis

| Coefficient estimates | Exponential values of the coefficient estimates | P value | Confidence interval |
|-----------------------|-----------------------------------------------|---------|---------------------|
| Fail                  | -1.806                                        | 0.16    | <0.0001             |
| Borderline            | -0.018                                        | 0.98    | 0.960               |

Legend: Coefficient estimates for ‘Fail’ score was -1.806 at statistically significant level (p < 0.0001).

For ‘Borderline’, the coefficient was -0.018 at statistically insignificant level (p = 0.960).

**Figures**
Overall experiment procedure Legend: UK OSCE examiners were recruited nationally and randomly allocated into two groups. They watched two mock OSCE videos in which a medical student played by an actor completed the task with and without the NNEA. The scores were collected online and statistically analysed.
Summary of the online platform pages Legend: Page 1 consisted of the introductory information for examiners, followed by page 2 which included the first video, supplementary information and the weblink to the online mark sheet. Page 3 contained second video and supplementary information.
Checklist sum dispersion for videos with and without the NNEA Legend: On the left is the box plot for the checklist sum given to videos with the NNEA (Q1 = 39, Median = 42, Q3 = 44.5, Max = 51, Min = 31). On the right is the box plot for the checklist sum given to videos without the NNEA (Q1 = 38, Median = 42, Q3 = 44.5, Max = 55, Min = 33).
Change in the checklist sum given by individual examiners in group 1 Legend: Points on the left are the sums given to the first video A (NNEA present, script 2) and points on the right are the sums given to the second video B (NNEA absent, script 1). The lines between the two point shows the changing trend for each examiner.
Figure 5

Change in the checklist sum given by individual examiners in group 2 Legend: Points on the left are the sums given to the first video C (NNEA present, script 1) and points on the right are the sums given to the second video D (NNEA absent, script 2). The lines between the two point shows the changing trend for each examiner.
Figure 6

Bar charts showing global scores for videos with and without the NNEA Legend: The bar chart on the left shows the number of global scores given to videos with the NNEA. The bar chart on the right shows the number of global scores given to videos without the NNEA. The counts for each grade groups are shown on top of each bar.
Change in the global scores given by individual examiners. Legend: The points on the left is the global scores given to the first videos A and C (NNEA present) and the points on the right is the global scores given to the second videos B and D (NNEA absent). The line between the two points express the trend in the change of scores. The thickness of the line and the number under each line with the matching colour represent the number of examiners who scored in that pattern.

**Supplementary Files**
This is a list of supplementary files associated with this preprint. Click to download.
Supplement1CONSORTMEED2000449.doc