BMJ Open Using the Many-Facet Rasch Model to analyse and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design

Mohsen Tavakol,[©] Gill Pinner

To cite: Tavakol M, Pinner G. Using the Many-Facet Rasch Model to analyse and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design. *BMJ Open* 2019;**9**:e029208. doi:10.1136/bmjopen-2019-029208

► Prepublication history for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-029208>).

Received 17 January 2019

Revised 24 July 2019

Accepted 24 July 2019

© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Medical Education Centre, The University of Nottingham, Nottingham, UK

Correspondence to

Dr Mohsen Tavakol;
mohsen.tavakol@nottingham.ac.uk

ABSTRACT

Objectives Sources of bias, such as the examiners, domains and stations, can influence the student marks in objective structured clinical examination (OSCE). This study describes the extent to which the facets modelled in an OSCE can contribute to scoring variance and how they fit into a Many-Facet Rasch Model (MFRM) of OSCE performance. A further objective is to identify the functioning of the rating scale used.

Design A non-experimental cross-sectional design.

Participants and settings An MFRM was used to identify sources of error (eg, examiner, domain and station), which may influence the student outcome. A 16-station OSCE was conducted for 329 final year medical students. Domain-based marking was applied, each station using a sample from eight defined domains across the whole OSCE. The domains were defined as follows: communication skills, professionalism, information gathering, information giving, clinical interpretation, procedure, diagnosis and management. The domains in each station were weighted to ensure proper attention to the construct of the individual station. Four facets were assessed: students, examiners, domains and stations.

Results The results suggest that the OSCE data fit the model, confirming that an MFRM approach was appropriate to use. The variable map allows a comparison with and between the facets of students, examiners, domains and stations and the 5-point score for each domain with each station as they are calibrated to the same scale. Fit statistics showed that the domains map well to the performance of the examiners. No statistically significant difference between examiner sensitivity (3.85 logits) was found. However, the results did suggest examiners were lenient and that some behaved inconsistently. The results also suggest that the functioning of response categories on the 5-point rating scale need further examination and optimisation.

Conclusions The results of the study have important implications for examiner monitoring and training activities, to aid assessment improvement.

INTRODUCTION

Measuring student's performance using an objective structured clinical examination

Strengths and limitations of this study

- The Many-Facet Rasch Model (MFRM) is well suited for analysing data from performance assessments, which are judged by human examiners.
- The results of the MFRM approach provide useful information about the quality of facets, that is, student's proficiency, the performance of individual examiners, the effectiveness of stations and domains.
- Fit statistics provide evidence of the examiner severity or leniency, which enable us to minimise the potential bias in severity calibrations.
- The MFRM provides useful feedback for examiners to help improve the consistency of their ratings.
- Several other factors can influence the measured performance of students, for example, circuits, the examination day and session (am, pm). We examined just four factors, we considered most important using this method.

(OSCE) provides important information for assessment leads beyond the summative results of simply pass or fail. OSCE results can be fed back to students in order for them to improve their learning; help faculty design and deliver the curriculum; ensure learning outcomes are covered and taught appropriately; provide information to improve future examination design and delivery to better evaluate student performance of learning outcomes; and provide decisive information for examiners, in order to learn from and adjust their ratings in the future. OSCEs must provide valid and reliable results to ensure that this wealth of information can be used appropriately by teachers, examiners, assessment setters and students in their decision-making about performance, subsequent attainment and progression. Unreliable results do not provide reliable feedback.¹ Additionally, an evaluation of student performance in an OSCEs should be independent of the day the

examination took place, the circuit and the session they were placed, and most importantly, the examiners that judged their performance. Similarly, performance ratings should be free from examiner conscious or unconscious bias and not depend on the student's gender, race and their perceived interest in the subject or previous work.

Examiner bias is the most common source of error in the rating process,² causing inconsistent behaviour across individual stations. Most OSCEs are designed relying on one examiner rating student performance on each station, but a number of independent examiners determining the overall mark, with an assumption that variability among individual examiners will even itself out across the examination. However, the pass/fail decisions for most OSCEs in medicine do not rely merely on the overall percentage mark but include specific criteria on numbers of stations to be passed to counter any 'compensation' for deficits in specific clinical skills. Examiner variability error distorts the domain being measured, leading to a threat to the validity and fairness of inferences drawn from performance assessments for specific tasks.³ However, it should be emphasised that any measurement is not immune from error. On evaluating the quality of any examination, assessment leads should pinpoint, isolate and estimate the sources of error in order to improve the utility of the examination and moderate if necessary.

A vast amount of psychometric research has been conducted outside the medical education arena in an attempt to examine the effect of various criteria (eg, examiner characteristics or examinee characteristics on performance ratings).⁴ However, most studies focus on the examiner effect in performance assessments where examiners observed and judged the performance of students, and the 'hawk' and 'dove' effect (or severity and generosity) can occur in performance ratings, particularly if different examiner teams rate students. When examiner bias is evident, students' scores may need to be moderated for differences between examiners assuming examiners have a common understanding of the performance being measured. Plausible approaches have been suggested to adjust examiner bias effects.^{5 6} A simple approach for detecting the examiner effects has been discussed elsewhere.⁷

In the field of medical education, studies have mostly been conducted using classical test theory (CTT), examples have included: evaluating the effect of 'extreme examiners',⁸ the effect of contrast error,⁹ examiner leniency at the start of an OSCE session¹⁰ and standard setting.¹¹ Although CTT is an often used and well-established model, item response theory (IRT) may offer some advantage. IRT focuses on item responses as unique rather than on whole tests. It focuses on a variety of metrics, not just raw score metric.¹² Using an IRT model, we find that a student's ability may differ from students who receive the same scores. Under the CTT model, students' test scores are grounded in a cohort of students (depending on the level of difficulty of the OSCE stations), rather than individual students. Consequently, they may not adequately

represent the quality of the OSCE stations themselves. It should be emphasised that students' test scores are not necessarily equivalent to their ability, because the ability is independent of the test and remains constant, unless new learning adds to student ability.¹³ Despite the fact that the IRT model can provide stable estimates of the features of students, examiners, domains and stations and how these features interact in describing performance rating and student ability, the IRT model for improving the quality of OSCEs is not widely used.

The Many-Facet Rasch Model (MFRM) is a classic method from IRT that can be used to investigate the effect of a number of relevant factors in an OSCE that may have an impact on student scores, notably the student themselves, the construct and subsequent domain content of a station and examiners. Below, we have attempted to present simply the key concepts and process required to interpret the outcomes of psychometric values in OSCE data using the application of the MFRM approach.

Theoretical framework: MFRM

Having highlighted that the CTT model focuses on the test and its errors but says little about how student ability interacts with the OSCE, its examiners, domains and stations, we suggest the MFRM as an alternative approach. The MFRM aims to measure the association between a student's ability and other factors such as examiners, domains and stations to improve OSCE quality.

For example, consider an OSCE examination where examiners measure the performance of students in agreement with a set of analytical criteria (eg, tasks/stations, domains). Here, relevant factors (usually termed 'facets') are student performance, the difficulty of the station/task, the behaviour of the examiner (eg, too harsh or too lenient), and the level of difficulty of the domains assessed. These facets are all likely to affect the scores awarded by the examiners both independently and synergistically. Examiners are a common source of unwanted bias and may generate important variations in student scores. Such variations can be a threat to the validity of the assessment outcomes of OSCEs.¹⁴ While there are many facets that potentially affect the scores awarded to students (eg, the circuit examined on, the day the examination is taken, the session, student characteristics such as gender), for the purpose of this study, we have selected the facets of 'students', 'examiners', 'domain' and 'stations' to be analysed. Any measurement may have many facets; the students are the object of measurement, so must be considered; the stations identifying the student ability for that particular station, with the domains identifying the ability of the student against the specific domain area tested are also key components of any assessment process, warranting investigation. Alongside these, the examiner facet is most likely to have the greatest impact, where examiner bias, variability, characteristics and behaviours are the most common source of error as previously described.²⁻⁴

MFRM is an extension of the Rasch model, which has been described elsewhere.¹⁵ In addition, interested

readers may refer to Bond and Fox, and Linacre^{16 17} for the technical aspects of the MFRM to obtain a greater understanding of the principles of parameter additivity and the symbols used for the Rasch family of models.

MFRM overcomes some limitations of the CTT model (eg, sample dependency) and allows us to estimate the probability that a student with a particular ability will produce a particular response to a set of facets (eg, examiners, domains and stations). For this study, where four facets are involved, we can estimate the probability of student N being given a rating of K by examiner X on domain Y and station Z. The ratings awarded for all domains and stations by examiners are used to estimate the performance of each examiner, the individual domain level of difficulty and individual station level of difficulty. Using the MFRM approach, we can gain useful diagnostic information about individual student performance, the behaviour of each examiner, the utility of each domain and station scored and the functioning of the response or rating scale, categories. Such analysis provides invaluable information for assessment leads to use to improve future OSCE design and implementations.

Response or rating scale categories: typically in OSCEs, students receive numerical ratings from examiners based on agreement categories or anchor descriptors found in a checklist or domain-based marking scoresheet. The categories are ranged from 0 to 4. The difference in rating between 2 and 3 is not equivalent to the difference in rating between 3 and 4. However, such rating scales yield ordinal scale (non-linear ordinal scale), not an interval/linear scale, especially at the end of categories where they denote infinite range of the performance of interest, for example, 'unsatisfactory' at the low end of the category and 'excellent' at the high end.¹⁷ To overcome this issue, the observed nonlinear ordinal ratings, using the Rasch model, are transformed into a continual, linear measure or the logit scale (or log-odds unit).¹⁵ Therefore, the elements of student performances, examiners, domains and stations are simultaneously calibrated onto the logit scale, and then they are placed on a common scale to compare within and between facets.

It is important to emphasise at this point that the available data may not necessarily fit the model, but by using fit statistics, we can explore the degree to which our data fit the model and determine whether this method of analysis is viable or not, before proceeding further.

Fit to the model: the analysis of fit shows how well the data generated from an OSCE examination fit the MFRM, that is, how well student ability, station difficulty, domain difficulty or examiner performance contributes to the domain of measurement. Fit statistics allow us to monitor student performance, examiner performance, domain and stations. The fit statistic is a quality control tool for managing and validating the facets of interest (in this study, the domains, examiner ratings and stations).¹⁸ Fit statistics, for each element within each facet, show the disparity between observed ratings and expected ratings given the student ability using a X^2 test or t-test.

Fit statistics show how well each examiner's ratings are consistent over students, stations and domains. For example, severe and lenient examiners will show unexpected performance ratings on students. If such irregularities are seen, these will require closer examination and may identify examiner leniency/stringency effects and/or other examiner effects such as central tendency or halo effect. Additionally, the fit statistics produced for each examiner provide useful information for monitoring and training of examiners. A simple approach to assess the overall data-model fit is to calculate the absolute standardised residuals. If 5% or less of the standardised residuals with absolute values are 'outside ± 2 ', the data will fit the model satisfactorily.¹⁹

Infit and Outfit statistics: within fit analysis, there are two fit statistics, infit and outfit statistics. Procedures for calculating fit statistics have been explained elsewhere.^{18 20} Conceptually, the infit statistic is the difference between observed ratings and expected ratings. A large difference between observed and expected can threaten the accuracy of the measurement. The infit statistic provides useful information about the association between student ability and other facets as they are sensitive to unexpected ratings and inlying observations. Put in another way, the infit statistic is sensitive to inliers (inlier sensitive), and an inlier rating is a rating to a facet whose difficulty is close to the student's performance, sometimes called the targeted observation. Unlike the infit statistic, the outfit statistic is sensitive to outlying deviations from the expected ratings (eg, when examiner leniency/stringency effects occur, or tasks are very easy or very hard).²¹ The expected values of these statistics are 1, indicating a perfect fit. Values between 0.50 and 1.50 indicate an acceptable fit. However, values less than 0.50 and greater than 1.50 are termed misfitting and overfitting, respectively.²² Fit statistics greater than 1.50 provide a distortion of measurement, values less than 0.5 indicate that the examiners were unable to discriminate high and low performers on tasks and domains leading to a misleading picture of the reliability of OSCE.¹⁷ An MFRM analysis also provides other metrics for monitoring and moderating the performance ratings awarded to students, such as a variable map, fair and observed average, reliability and separation, SEs and the functioning of the rating scale. The variable map is a very important output of the MFRM and provides a comparison of all facets to each other on a common scale.

Research questions

This research seeks to address the following questions:

1. To what extent do the facets modelled in the OSCE administration (students, domain, station and examiners) contribute to scoring variance, and fit into an MFRM of OSCE performance?
2. Do the rating scales function, ranging from 0 and 4, well in estimating the domains in question?

METHOD

Students and OSCE procedure

A 16-station OSCE was conducted to 329 final year medical students (186 females and 143 males) at the University of Nottingham. This is the final clinical examination for the medical course curriculum covering learning outcomes mapped to General Medical Council (GMC) requirements documented in 'Outcomes for Graduates'. For logistical reasons, the examination was divided into two sets of eight stations, each OSCEs being held over 3 days with multiple circuits happening simultaneously and a number of cycles across each day. This allowed a large number of students to be put through the examination at a single site. The OSCE was marked using specialist software on Android tablets, the data collected online and then analysed. All OSCE stations were of 10 min duration with 2 min for moving and reading the next station instructions. Each station was scored using a domain-based marking scheme; eight fixed domains were used across the whole OSCE, a single station sampling 4 or 5 of these domains and the examination blueprinted to ensure appropriate coverage of domains and curriculum learning outcomes. The eight domains used for all clinically based examinations are communication skills, professionalism, information gathering, information giving, examining, clinical interpretation, diagnosis and management. Communication skills and professionalism are assessed in every station, and all domains are weighted to ensure appropriate attention to the domain of the individual station, and a composite score is derived. To accomplish this aim, a panel of clinicians judge and assign a weighting for each domain for every station.

Domains were rated according to the detailed anchor descriptors of the performance on the domain of interest. Based on these anchor descriptors, examiners provided final ratings for each domain, as well as provide a global rating score for the purpose of standard setting. On each domain, student performance was rated using a 5-point scale for domains (except the procedure domain where an embedded checklist using a binary scale, that is, 0 and 1 is used). The scores range from 0 (the student does not cover the areas within the domain) to 4 (the student covers areas within the domain accurately and comprehensively). The OSCE was found to have satisfactory reliability with a G-coefficient of 0.71.

Examiners

The examiners who scored students' performance at each station were all clinicians, mostly doctors, but included a small number of specialist health professionals involved in teaching students, examining within their fields of expertise. Examiners attended half-day face-to-face training prior to the examination, as well as a video of the specific station they were to examine a few days prior to the assessment so that they could be fully familiar with the topic and what to expect. They were given specific guidance on the range and criteria of marking with station-specific anchor descriptors. One examiner rated each student

within each OSCE station, a total of 172 examiners participated in the scoring of students' clinical performance.

Multifaceted data analysis

The OSCE data were analysed using FACETS, a computer program for Rasch analysis of examinations with more than two facets.²³ OSCEs are not immune from error due to different facets (eg, examiners, cases, domain or stations). The MFRM approach enables us to examine the trustworthiness of the OSCE data and student data-model fit. Therefore, the ratings that examiners awarded to students were used to estimate individual student performance domain (domain) difficulties, the severity/leniency of the examiners across all the domains and station difficulties. Within the MFRM approach, the students, examiners, domains, stations and rating scales are calibrated onto the same equal-interval scale, named 'logits'. The MFRM estimates a measure for each element (student, examiner, domain and station) of each facet (students, examiners, domains and stations) located in a single frame of reference for interpreting the FACETS output. Increasing positive logit values imply greater levels of student ability, examiner stringency and higher levels of domain and station difficulty. Negative logit values indicate lower levels of student ability, examiner leniency and lower levels of domain and station difficulty.

Each element of each facet was calculated using a logit measure and its SE, as well as the analysis of model fit statistics (infit statistics and outfit statistics). Fit statistics refer to the degree to which the observed ratings of each element of each facet fit with the expected ratings of each element of each facet that are produced by the MFRM. An observed rating is the rating awarded to students on a specific domain. An expected rating is the rating that the examiner will award to students on a specific domain, given the characteristics of the examiner (ie, stringent or lenient), other examiners' ratings of the student on that domain, the estimated difficulty level of the domain and the estimated student performance. For examiners, a large difference between the observed and expected ratings is a potential threat to the reliability of the ratings awarded to students.²⁴

The separation reliability statistic is equivalent to Cronbach's alpha, ranging from 0 to 1. Therefore, a low separation of examiner reliability is desirable as we do not want to view the variability across examiner ratings. The separation reliability statistic is calculated for other facets, that is, student, domain and station.

Single examiner-rest of the examiner correlation (SE-ROE) is also reported to identify the degree to which the behaviour of each examiner is consistent with the behaviour of the other examiners. For example, if an examiner significantly had a lower SE-ROE correlation than other examiners, this would suggest that the examiner has not behaved consistently and that she or he has awarded ratings more randomly.

To assess the quality of the rating scales for the domains of interest, the functioning of the rating scale categories

was investigated. The rating scales categories provided for each domain can affect the reliability of OSCE scores. Therefore, examiners should consider these categories are important equally and require equal attention. At least 10 observations are required for each category. Erratic observations across categories may be evidence of implausible category usage. Therefore, a uniform distribution of observations across categories is a desirable outcome. In addition, the average measures of each category 'must advance monotonically up the rating scale' with an outfit mean square less than 2.²⁵

Patient and public involvement

In this study, patients and the public were not involved.

RESULTS

Model-data fit

Psychometrically speaking, the Rasch model is based on the empirical data, in this instance, the OSCE data. No empirical data will fit the Rasch model perfectly,¹⁷ and one must explore the aberrant observations for examiners, domains and stations to improve the fit of the examination data to the model, and to improve the development of OSCEs. Fit statistics for each facet will be addressed later.

In this study, the valid responses used for the estimation of model parameters is 32 242. Of these, 1241 responses (3.8%) have a standardised residual outside of ± 2 , suggesting the OSCE data fit the model and MFRM approach is appropriate to use.

Variable map

Figure 1 shows the variable map. It presents the spread of students' ability, examiner severity measures, domain difficulties and station difficulties on the same logit scale (the first column, titled 'Measr'). Therefore, all facets can be compared with each other on a common scale. The second column (titled 'student') presents student ability. In this column, each dot represents one student, with six students represented by an asterisk (*). As the 'average' student is set at the zero on the logit scale, a high positive logit indicates a high level of ability, and a low negative logit indicates a low level of ability. The third column shows the examiner facet. Stringent examiners are located at the top of the column, and lenient examiners locate at the bottom of the column. The domain facet appears in the fourth column and indicates each domain's average level of difficulty 16 stations on the logit scale. The domains are distributed from most difficult at the top to the least difficult at the bottom. The fifth column contains the station facet. The sixth column represents 'total score achieved' and the seventh column shows the 'domain rating scale' (0–4). The stations are distributed from most difficult at the top to the least difficult at the bottom.

Student facet: The second column on the variable map shows student ability and suggests that the students are 'cleverer' than the domains, with the exception of the 'Management' domain for station 16 (S16Mang). However, most stations map into student ability.

Figure 1 Variable map from the Many-Facet Rasch measurement. The forth column headed 'domain' displays each of the 8 scoring domains occurring across the 16 stations examined. Comm, communication; CR, clinical reasoning; Exam, examination; InfGath, information gathering; InfGiv, information giving; Mang, management; Prof, professionalism; Proc, procedure; Stations, S1–16.

Table 1 Calibration of the domain facet

Domain	Measure	SE	Infit	Outfit	N
Professionalism	-0.83	0.10	0.83	0.8	16
Procedure	-0.58	0.09	1.44	1.43	4
Communication	-0.43	0.09	0.77	0.76	16
Clinical reasoning	0.08	0.09	1.12	1.09	12
Information gathering	0.32	0.08	0.92	0.92	9
Information giving	0.37	0.08	0.83	0.83	6
Examination	0.59	0.08	0.97	0.97	8
Management	1.07	0.08	0.97	0.96	11

Examiner facet: Figure 1 also shows that examiners are relatively lenient. Examiners 84 and 157 were most severe with logits of 2.10 and 1.99, respectively, and examiner 148 was the most lenient (-1.75 logits). As we can see from the variable map, the examiners exhibit a difference in sensitivity of 3.85 logits, though the difference is not statistically significant if we assume that the examiners come from a normal distribution, $\chi^2_{(170)}=167.6$, $p=0.54$. The examiner separation reliability of 0.98 indicates that the examiners are not similar in their ratings. The variable map in figure 1 supports this finding, as the examiners are not located at the same logit level.

An examination of the infit and outfit statistics reveals the highest and the lowest values of infit and outfit statistics in some examiners. The infit statistics ranged from 0.27 to 4.59 logits and had a mean (SD) of 1.71 (0.41). The outfit statistics ranged from 0.27 to 2.45 logits and had a mean (SD) of 0.99 (0.25). These high low values suggest that the ratings awarded by the examiners were more erratic than expected by the model in terms of the

rated domain difficulty. In other words, these examiners might give high ratings within a difficult domain or to low performers or a low rating within an easy domain or to high performers. Such findings can provide useful feedback for examiners in order for them to improve their consistency in ratings and ultimately making students' marks fair.

Further analysis of SE-ROE correlation shows that the correlation of each examiner with others ranged between 0.14 and 0.92, indicating that some examiners have not behaved in a consistent manner. To ensure the examiner severity/leniency, the SE-ROE correlation should be reviewed for each examiner.

Domain facet: The calibration of the domain facet is addressed in table 1. The difficulty levels of the domains range from 1.07 logits (SE=0.08) for 'Management' (most difficult) to -0.83 logits (SE=0.10) for 'Professionalism' (least difficult).

Figure 2 shows the item characteristics curve for each domain. For the domain of 'Professionalism', a student with the average ability (0 logits) has approximately 72% chance of performing this domain correctly. However, for the domain 'Management', the same student has approximately 28% chance of performing this domain correctly. The overall differences between the domains are significant, $\chi^2_{(81)}=5345.1$, $p=0.00$, with high reliability of separation index of 0.99. There is no evidence that the examiners behaved inconsistently, and all of the fit statistics are between 0.77 and 1.44. As shown in the variable map, figure 1, the domains are well targeted at the cohort of the examiners in terms of difficulty, suggesting the anchor descriptors are mostly well addressed.

Station facet: The last facet examined is the OSCE station itself. As is evident from figure 1, station 8 is the

Figure 2 Item characteristic curves for eight domains.

Table 2 Calibration of stations

Domain	Measure	SE	Infit	Outfit
Station 1	2.62	0.03	1.50	1.51
Station 2	1.04	0.03	2.24	2.25
Station 3	3.32	0.03	1.10	1.10
Station 4	3.25	0.03	1.03	1.04
Station 5	2.23	0.03	1.58	1.54
Station 6	4.27	0.03	1.22	1.22
Station 7	3.15	0.03	1.35	1.36
Station 8	4.44	0.03	0.90	0.90
Station 9	3.14	0.03	1.39	1.39
Station 10	3.40	0.03	1.18	1.17
Station 11	2.22	0.03	1.49	1.49
Station 12	3.29	0.03	1.19	1.17
Station 13	3.27	0.03	1.15	1.15
Station 14	2.36	0.03	1.82	1.81
Station 15	3.10	0.03	1.28	1.29
Station 16	1.24	0.03	1.01	0.99

most difficult station (4.44 logits) and stations 2 and 16 were the least difficult (1.04 and 1.24 logits, respectively) according to the ability of students. In terms of difficulty measures, the stations ranged from 1.04 to 4.44 logits, indicating a relatively large spread (3.40), which is supported by the station separation index, a value of 0.99. When constructing the 16-station OSCE and determining the variety of station tasks to be included, the examination is designed to sample a range of skills, which is required in real-life clinical practice and the examination is, therefore, expected to have a range of the station difficulty. This is borne out by the results which show that the stations have a range of level of difficulty, which is statistically significant, $\chi^2_{(15)}=11.3810.1$, $p=0.00$. The station difficulties, SEs and infit, and outfit statistics are presented in table 2. The larger the logit measures, the less likely students will fail the station (depending on the credibility of the pass mark). All stations showed a good fit with both infit and outfit statistics close to 1 except stations 2, 5 and 14.

The functioning of the domain rating scale

To address how well the examiners interpret the rating scale categories for each domain across all stations, the

functioning of the assigned rating categories was examined. The purpose was to empirically identify the optimal number of rating scale categories for measuring the domains. Table 3 shows the category functioning for each response option. On the basis of this table, the category frequencies (total) are negatively skewed with at least 10 observations in each category. However, the averages measures are disordered, and they are not increased monotonically across the rating scales. The issue is category 1, with an average mean of -0.30 logit. An examination of fit statistics for assessing the quality of rating scales shows that category 0 has an outfit mean square greater than other categories, suggesting introducing noise into the measurement process. These results may suggest combining a five-category rating scale into a four-category rating scale.

DISCUSSION

There is a vibrant and growing array of psychometric methods, which allow the interrogation of data from assessments. These methods can inform assessment faculty such that they are better placed to make reasoned judgements and decisions as to the outcome for individual students. These decisions, however, are only based on how a student performs in front of assessors and does not necessarily predict how they will perform in the real-world clinical environment. As an assessment tool, OSCEs are now widely used in undergraduate and postgraduate medicine globally. OSCEs have become popular mostly due to their perceived intention of being objective rather than subjective in nature and can be standardised, thus being more 'fair' to candidates. This allows a student mark to be calculated, a score generated and a pass/fail decision determined. However, clinical assessments, including OSCEs, are not really about assigning marks.²⁶ They are intended to help faculty make a reasoned decision whether or not a student has attained a level of clinical proficiency adequate to allow safe practice at a particular level of study from what they have observed.

The OSCE may investigate the students level of ability to perform certain element (or domain) of a task designed to measure that element or domain, for example, professionalism. If a student passes the professionalism domain, does that necessarily mean the student is proficient in that domain of interest, that is, is professional? Rasch²⁷ alerted that 'Even if we know a person to be very capable, we cannot be sure that he will solve a difficult problem,

Table 3 Summary of the functioning of the domain rating scale

Category	Total	Percentage	Average measure	Expected means	Outfit
0	37	0	1.27	-2.82	3.3
1	652	2	-0.3	-0.1	1
2	3086	11	2.26	2.31	0.9
3	11982	44	2.96	3.01	0.8
4	11221	42	3.81	3.74	0.9

nor even a much easier one. There is always the possibility that he fails—he may be tired, or his attention led astray, or some other excuse may be given' (p.73).

Having emphasised that assessments, including OSCEs, are prone to both external and internal errors such as the examiner's judgement, the students' ability and the imperfect nature of the measurement in our assessment, it could be argued that too much reliance is placed on the student performance rating at an OSCE in determining a student's capability when they face real-world situations. Alternatively stated, student marks are not equivalent to evidence for overall proficiency in the domains of interest. Having emphasised that multiple facets such as examiner severity or leniency, station difficulty and domain difficulty can influence the quality of judgements about a student's ability, this can lead to unfair, unreliable or incorrect inferences of the student's proficiency.

As already stated, examiner error is an important reason for an incorrect inference or bias within an OSCE (ie, generosity error, severity error, halo error and central tendency error) and it is, therefore, necessary to detect erratic examiners for potential biases or errors, commonly severity bias. In this study, the MFRM allows us to detect erratic examiners and to collaborate other the facets of interest, the student performance, the domain facet and the station facet.

The analysis of our data suggests that examiners did not behave consistently across stations, although the difference between them is not statistically significant. Furthermore, an inspection of the relationship between student and examiner, using the variable map (column 2 and column 3), shows that the ability of students does not map well with the performance ratings awarded by the examiners. Based on our previous psychometric analyses from OSCEs and also knowledge-based tests, we have seen similar distributions, that is, assessments are relatively easy, and examiners are relatively lenient. This could be due to the fact that the questions or stations selected should give students 60%–80% chance of success, and so that examiners are usually located one or more logit below the ability of students.²⁸ In addition, undergraduate medicine examinations are intended to fail only students with a below minimum benchmark ability, so if OSCEs or knowledge-based tests are too difficult for students, most of the students will fail the assessment, which is not the purpose of these assessments.

An important finding from this study is that the domains fit the MFRM model with high reliability of separation index. Additionally, the domains are well mapped with the performance of the examiners. However, the ability of students lies beyond the difficulty of domains, except for the management domain in station 16. Further analysis of domains shows that the most challenging domain relative to others assessed was 'Management', and the least difficult domain was 'Professionalism'. This is not unexpected for an undergraduate final year examination. To score a low mark on professionalism, students would need to demonstrate consistently poor behaviours and values that would

be unusual as they are highly aware that they are being observed. Because of this in our OSCEs, the professionalism domain has a lower weight compared with other domains.

At undergraduate level, we may expect students to find formulating a management plan most challenging, as it is arguably the most advanced area of practice for them, and the one least practised. However, to our knowledge, there is no evidence either consistent or inconsistent with these findings from other sources. Further work, therefore, is required to establish these consistency of these results.

When exploring the difficulty hierarchy of stations (figure 1), most stations match well with student ability, which is encouraging. However, further analysis shows that some stations do not fit the MFRM (table 2), and hence, these stations need reviewing in order to improve the quality of stations.

This study also set out with the aim of assessing the functioning of the rating scale of the domains. The results suggest that a four-category rating scale would be better than the current five-category scale for each domain as the first category is largely redundant (score of 0). However, this would not be consistent with results from other studies which suggest that higher reliabilities are achieved through using 5-point scales or more.^{29 30} A study by Preston and Colman did suggest contrary to that previously stated, that the acceptance of 5-point scales is less justified and that discriminating power is low for scales with just 2, 3 or 4 points. Discriminating power is high for scales with 9 or 10 response modes at a level of statistical significance.³¹ However, a review of the literature provided support for scales consisting of 7 points,³² which supports Miller's idea of the capacity for processing information and 'the span of immediate memory'.³³ Caution should, however, be exercised in generalising the optimal number of response modes in rating scale functioning of domains because examiners awarded the ratings via observations, not self-administered questionnaires, which measure attitudes or satisfaction for service quality and improvement. Therefore, further research studies, based on objective rating scales, are required. A further limitation that only University of Nottingham, final year medical student OSCE data was interrogated, may limit the generalisability of the results to inform other medical schools or indeed other healthcare courses.

CONCLUSIONS

OSCEs are very dependent on the quality ratings awarded by examiners. About 85 years ago, Guilford stated 'Raters are human, and they are therefore subject to all the errors to which humankind must plead guilty'.³⁴ In OSCEs, examiner errors or bias are an essential source of domain-irrelevant variance, and therefore, threats the validity of the OSCE scores at the individual level and group level.³⁵ Consequently, it is important that errant examiners are identified and trained in high-stakes assessment because they can affect the validity, reliability and fairness of the OSCE scores.³⁶ The MFRM presents a promising method in the monitoring of examiner stability. An MFRM analysis

enables assessment leads to detect and measure the performance of examiners. Finally, the moderation committee can monitor the examiners' scoring behaviour, based on the results of fit statistics, in order to correct student marks and make a fair OSCE and as free as possible from sources of bias and error, for example, examiner error. In addition, providing feedback on the ratings awarded enables examiners to understand entirely the impact of what is they are performing so that they can improve their ratings. The results of MFRM provide invaluable information for gathering this kind of individualised feedback, especially when individualised feedback is combined with face-to-face training as this will help examiners to pay attention to individual biases.³⁷

The results of an MFRM also enable assessment leads to detect the infit of students, domains and stations in order to improve the quality of checklists. Furthermore, an analysis of the functioning of response categories for each domain will enable assessment leads to improve the validity and reliability of OSCEs by choosing an optimal number of response categories in rating scales. Therefore, as one of the aims of examiner training is that the examiners have a common understanding of response categories, presentation and discussion of these results should be included in examiner training. Taken together, this study has shown that the MFRM is a powerful measurement approach, and solves issues attached to measurement in assessments and hence results in the improvement our assessments, especially where there is some uncertainty about examiner behaviour.

Contributors MT and GP developed the conception and study design. MT collected and analysed the data. Both contributed to the methodology and interpretation of the data. MT initially drafted the paper. GP edited and revised the manuscript. Both authors approved the final manuscript before submission.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

REFERENCES

1. Brookbart S. *The art and science of classroom assessment: the missing part of pedagogy*. Washington, DC: The George Washington University, Graduate School of Education and Human Development, 1999.
2. Cronbach L. *Essentials of psychological testing*. New York: Harper and Row, 1990.
3. Lane S, Stone CA, assessment P. Performance assessment. In: Brennan RL, ed. *Educational measurement*. Westport, CT: Prager, 2006: 387–431.
4. Landy FJ, Farr JL, rating P. Performance rating.. *Psychol Bull* 1980;87:72–107.
5. Houston WM, Raymond MR, Svec JC. Adjustments for rater effects in performance assessment. *Appl Psychol Meas* 1991;15:409–21.
6. Paul SR. Bayesian methods for calibration of examiners. *Br J Math Stat Psychol* 1981;34:213–23.
7. Tavakol M, Pinner G. Enhancing objective structured clinical examinations through visualisation of checklist scores and global rating scale. *International Journal of Medical Education* 2018;9:130–4.
8. Fuller R, Homer M, Pell G, et al. Managing extremes of assessor judgment within the OSCE. *Med Teach* 2017;39:58–66.
9. Yeates P, Moreau M, Eva K. Are examiners' judgments in OSCE-Style assessments influenced by contrast effects? *Acad Med* 2015;90:975–80.
10. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. *Med Teach* 2015;37:81–5.
11. Margolis M, Clauser B. The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice* 2014;33:15–21.
12. Yen W, Fitzpatrick A. Item response theory. In: Brennan RL, ed. *Educational measurement*. Westport, CT: American Council on Education, 2006: 111–53.
13. Hambleton R, Jones R. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: issues and practice* 1993;12:38–47.
14. Eckes T. *Introduction to Many-Facet Rasch measurement*. Frankfurt: Peter Lang AG, 2015.
15. Tavakol M, Dennick R. Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide No. 72. *Med Teach* 2013;35:e838–48.
16. Bond T, Fox C. *Applying the Rasch model: fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates, Publishers, 2007.
17. Linacre J. *Many-faceted Rasch measurement*. Chicago: MESA Press, 1994.
18. Wright B, Stone M. *Measurement essential Delaware*. Wide Ran, Inc, 1999.
19. Linacre J. Unexpected (standardized residuals reported if not less than) = 3. Available: <https://www.winsteps.com/facetman/unexpected.htm>
20. Wright B, Stone M. *Best test design*. Chicago: MESA, 1979.
21. Wright B, Masters G. *Rating scale analysis*. Chicago: MESA Press, 1982.
22. Linacre J. What do Infit and outfit, mean-square and standardized mean? *Rasch Meas Trans* 2002;16.
23. Linacre J. *Facets, a computer program for the analysis of multifaceted data*. Chicago: MES Press, 2018.
24. Myford C. Detecting and measuring rater effects using many-facet rasch measurement : Part II. *Journal of applied measurement* 2003;4:386–422.
25. Linacre J. Investigating rating scale category utility. *Journal of Outcome Measurement* 1999;3:103–22.
26. Levy R, Mislevy R, Modeling BP, et al. *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press, 2016.
27. Rasch G. *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press, 1980.
28. Linacre J. Computer-adaptive testing: a methodology whose time has come, 2000. Available: <https://www.rasch.org/memo69.htm>
29. Jenkins GD, Taber TD. A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology* 1977;62:392–8.
30. McKelvie SJ. Graphic rating scales - How many categories? *Br J Psychol* 1978;69:185–202.
31. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 2000;104:1–15.
32. Cox EP. The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research* 1980;17:407–22.
33. GA M. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 1956;101:343–52.
34. Guilford J. *Psychometric methods*. New York: McGraw-Hill, 1936.
35. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 1995;50:741–9.
36. American Educational Research Association APA, National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, 2004.
37. Knoch U, Read J, von Randow J. Re-training writing raters online: how does it compare with face-to-face training? *Assessing Writing* 2007;12:26–43.