Rasch analysis on OSCE data : An illustrative example

Tor E, Steketee C

School of Medicine, The University of Notre Dame Australia, Fremantle

RESEARCH

Please cite this paper as: Tor E, Steketee C. Rasch analysis on OSCE data: an illustrative example. AMJ 2011, 4, 6, 339-345
http://doi.org/10.21767/AMJ.2011.755

Corresponding Author:
Elina Tor
School of Medicine, Fremantle
The University of Notre Dame Australia
45 Henry St (PO Box 1225),
Fremantle WA 6959
Email: elina.tor@nd.edu.au

Abstract

Background
The Objective Structured Clinical Examination (OSCE) is a widely used tool for the assessment of clinical competence in health professional education. The goal of the OSCE is to make reproducible decisions on pass/fail status as well as students' levels of clinical competence according to their demonstrated abilities based on the scores. This paper explores the use of the polytomous Rasch model in evaluating the psychometric properties of OSCE scores through a case study.

Method
The authors analysed an OSCE data set (comprised of 11 stations) for 80 fourth year medical students based on the polytomous Rasch model in an effort to answer two research questions: (1) Do the clinical tasks assessed in the 11 OSCE stations map on to a common underlying construct, namely clinical competence? (2) What other insights can Rasch analysis offer in terms of scaling, item analysis and instrument validation over and above the conventional analysis based on classical test theory?

Results
The OSCE data set has demonstrated a sufficient degree of fit to the Rasch model ($\chi 2 = 17.060$, DF=22, p=0.76) indicating that the 11 OSCE station scores have sufficient psychometric properties to form a measure for a common underlying construct, i.e. clinical competence. Individual OSCE station scores with good fit to the Rasch model ($p > 0.1$ for all $\chi 2$ statistics) further corroborated the characteristic of unidimensionality of the OSCE scale for clinical competence. A Person Separation Index (PSI) of 0.704 indicates sufficient level of reliability for the OSCE scores. Other useful findings from the Rasch analysis that provide insights, over and above the analysis based on classical test theory, are also exemplified using the data set.

Conclusion
The polytomous Rasch model provides a useful and supplementary approach to the calibration and analysis of OSCE examination data.

Key Words
Medical education, clinical skills assessment, OSCE, Rasch analysis

What this study adds:

1. This study exemplifies the potential insights from Rasch analysis for clinical teachers and other stakeholders through a retrospective analysis of real OSCE data.
2. There are no published reports in the literature on the use of polytomous Rasch modelling as a quality assurance tool for OSCE data. This study provides concrete examples to demonstrate the practicality of Rasch analysis for OSCE data.
3. Implications for future development in medical education assessments are discussed.

Background

In assessing clinical competence of undergraduate students, medical schools typically use the OSCE. The goal is to make reproducible decisions on the pass/fail status and students'

clinical competence according to their demonstrated abilities based on the OSCE scores. Therefore, medical schools must establish empirical as well as conceptual evidence of validity and reliability of OSCE scores, to indicate that these scores are true measures of students' clinical competence. Traditionally, medical schools adopt Classical Test Theory (CTT) where raw scores are taken as measures for students' clinical competence. Reliability and validity of OSCEs are also evaluated using raw scores. However, an emerging measurement paradigm based on the work of Georg Rasch, a Danish mathematician, promises a powerful method for interrogating clinical assessment data, resulting in more valid measures for students' clinical competence to inform defensible and fair decisions on students' progression and certification. For example, the multifacet Rasch model estimates students' true measures of clinical competence by partitioning the variance in raw scores into variance due to item difficulty, student ability and examiner severity/leniency.[1, 2] This method, however, is rarely applicable in OSCE examination data in most medical schools, as it is reliant upon a crossed-design where all examiners must examine all students and all stations, at one point or another, throughout the OSCE. Given the paucity of examiners in medical schools, this is not a practical expectation.

Nevertheless, the Polytomous Rasch Model (PRM), an extension of the dichotomous Rasch model which is widely used to evaluate Multiple Choice Questions (MCQs) and Extended Matching Questions (EMQs) examination data, [3] can yield the same rich outcomes when used in conjunction with CTT. There are no published reports in the literature of it being used to evaluate OSCE data, possibly due to the fact that this method only accounts for two components of variance in the analysis of assessment data, i.e. student ability and item difficulty. We would like to illustrate in this paper how the PRM can be used to evaluate the psychometric properties of the OSCE data for quality assurance with the same rigour.

Data modelling based on the Rasch model

As implied by the term Item Response Theory (IRT) where Rasch models belong, Rasch modelling provides a scaling method to establish measurement (measures) based on students' response pattern to test items. The Rasch models are a class of their own in IRT. Rasch models are the only group of IRT models that have the criteria of 'objective measurement' such as unidimensionality, local independence and sufficiency embedded in its mathematical formulations.[3] The fundamental measurement paradigm in Rasch modelling is that to be qualified as an 'objective measurement' for a target construct, the response pattern to test items for that construct should *approximate* the pattern expected by the Rasch model.[2] Rasch modelling operates as a quality assurance framework for measurement as elaborated elsewhere.[4,5] This is in contrast to other measurement models under IRT that seek to describe a data set, in that if a set of data does not fit a model, the researcher should look for other models that will accommodate a fit. [4, 5] Anomalies in the response pattern or misfit of data to the model will be flagged through multiple graphical and statistical indicators.[6] As such, post-hoc investigations by the researcher involved are critical in Rasch analysis. With sufficient conceptual and/or theoretical support of the target construct, Rasch analysis also allows and facilitates experimental removal of inconsistent responses. Rasch modelling and analysis of the OSCE data served to flag anomalies in the raw scores and facilitate the decision to exclude (or retain) individual OSCE station scores where the response pattern deviated from the pattern expected by the Rasch model.

In its simplest form, when a student is rated for their performance in a task, the log odds of a student being rated in category x over the previous category (x – 1) is modelled in the PRM as a function of the student's ability and the task difficulty: [7]

Log $\frac{P(X_{vm} = X)}{P(X_{vm} = X - 1)} = \theta_v - \delta_m$

$P(X_{vm} = X)$ is the probability of being rated in category x

$P(X_{vm} = X - 1)$ is the probability of being rated in category x-1

θ_v is the ability of student *v*

δ_m is the difficulty of task *m*

The linear (additive) formulation of this model enables the separation of parameter estimates. Therefore, task difficulty is estimated independent of the combination of students' performance in the task. Similarly, student ability estimation is done independent of the set of tasks performed. As a result of this criterion of invariant comparisons across students and tasks, the characteristic of *sufficiency* is achieved. In other words, when the data fits the PRM, total raw scores become the sufficient statistic, containing all information about item difficulty and students' abilities pertaining to the underlying construct measured by the items.[4]

Method

Data design and Structure of the OSCE

The 2009 OSCE data set for fourth year students comprised 10 stations of simulated and structured clinical scenarios, each targeting clinical skills across different medical sub-disciplines and one non-clinical scenario station for the evaluation of professional and clinical reflective practices. Table 1 outlines the nature of the 11 stations and the aspects of clinical competence they were designed to assess.

The 10 clinical OSCE stations were conducted in two sessions on the same day. The portfolio station was in the afternoon on a different day. In each session, 40 students were randomly assigned into four equal-sized groups of 10 students each. In each session, there were four examination rooms for each station to enable four students to be assessed for the same station concurrently. In each examination room (except station five and station 11 where no simulated patient was involved), the structured clinical scenario was simulated by a standardised patient or a real patient. A clinician examiner was stationed in each room to rate the student's performance in his/her encounter with the simulated or real patient. Therefore, there were four different examiners per session in four different rooms for each station. Each student rotated from one station to another in a circuit and completed all 10 stations in one session. For the afternoon session, a similar set-up was used. Only some of the examiners returned for the afternoon OSCE session. The number of examiners involved in each station is shown in Table 1.

The OSCE was designed based on the assumption that all stations are measuring a common underlying construct, i.e. the clinical competence of the students. Students' overall marks for the OSCE were derived from the average of their individual station marks. The overall marks were used to indicate the applied clinical competence across the range of medical / surgical disciplines which comprised the fourth year MBBS curriculum.

Station	Discipline /Topic	Target Competency	No. Examiners
1	Medicine – Pulmonary Fibrosis	Physical examination; Presenting findings; Differential diagnosis; Further investigation	5
2	General Surgery – Abdominal Pain	History; Examination; Diagnosis; Investigation; Management	7
3	Critical Care & O&G	History; Short term & long term management; Interpreting symptoms	5
4	Musculo-skeletal	Examination; Diagnosis; Interpretation	5
5	Medicine – Tiredness	Investigation; Interpreting results; Further history; Diagnosis, Management	5
6	Critical Care	Investigation; Diagnosis; Interpretation of results; Management	5

7	Surgery – Anaesthetic	Informed consent; Material risk; Patient communication	7
8	Psychiatry - Rural Practice	Observe & note symptoms and signs; Diagnosis; Investigations and interpretation of results; Management	6
9	ENT	Neck examination; Results interpretation; Diagnosis; Discussion of management plan.	6
10	Palliative Care	History & discussion of management plan	4
11	Portfolio	Life-long learning skills and reflective practice skills	9

Table 1: Structure of OSCE

Research questions & Data analysis

In an effort to determine whether the clinical tasks assessed in the 11 OSCE stations map on to a common continuum of clinical competence (i.e. unidimensional), and what the PRM could offer in scaling, item analysis and instrument validation over and above the conventional measurement based on CTT, each of the station scores were analysed as one item. The maximum mark for each station was 20. The raw scores for each station were first collapsed into 10 categories of zero to nine to be fitted to the PRM, using RUMM2020 software.

Results

Overall scale fit to model: The item-trait interaction fit statistics evaluate the *suitability* of the data to establish a construct and its measures.[8, 9] The target construct in the OSCE is students' clinical competence. Rasch analysis provides a formal test for the unidimensionality of clinical tasks assessed in all stations and therefore the validity of using the summed scores as measures for students' clinical competence.

A non-significant $\chi 2$ statistic in the test of fit for the OSCE data ($\chi 2 = 17.060$, DF=22, p=0.76) as highlighted in Figure 1 indicates the consistency of the common underlying construct across the stations. The global fit of the scale constructed from 11 station scores also indicates that the hierarchy of station difficulty is consistent across the various levels of students' clinical competence on the scale. This again implies that tasks assessed in all stations do map on to a common underlying construct. Therefore, it is justified and valid to take the summed scores across stations as the indicator of students' levels of clinical competence.

ITEM-PERSON INTERACTION

	ITEMS		PERSONS	
	Location	Fit Residual	Location	Fit Residual
Mean	0.000	0.224	1.064	-0.041
SD	0.591	0.478	0.519	0.746
Skewness		-0.596		-0.928
Kurtosis		-0.329		3.238
Correlation		-0.100		0.216

Complete data DF = 0.860

ITEM-TRAIT INTERACTION		RELIABILITY INDICES	
Total Item Chi Squ	17.06	Separation Index	0.704
Total Deg of Freedom	22.00	Cronbach Alpha	0.677
Total Chi Squ Prob	0.76		

LIKELIHOOD-RATIO TEST	POWER OF TEST-OF-FIT
Chi Square	Power is GOOD
Degrees of Freedom	[Based on SepIndex of 0.704]
Probability	

Figure 1: Summary test of fit statistics – overall OSCE exam

Individual station fit to model

Table 2 shows the statistical evidence for individual item fit to the PRM i.e. fit residual statistics in Column 4 and Chi Square Statistics in Column 5 to Column 8. The *p* value in Column 8 is a test for null hypotheses of no differences in the observed and the expected item location. A *p*-value of > 0.05 indicates that the evidence from this data set fails to reject the null hypotheses of no differences between the observed students' scores and the Rasch predicted students' scores. In other words, a *p*-value of 0.96 as recorded for station five indicates that 96% of the differences between the observed and the model predicted scores happened just by chance. Scores from all stations

have demonstrated good fit to the model as demonstrated in the fit residual statistics, which are all within the range of + / - 2.5 and *p*-values above 0.05. These empirical findings indicate that each station is measuring a common underlying construct.

Station	Location	SE	Residual	DF	ChiSq	DF	Prob (*p*)
Station 9	-1.37	0.14	0.94	68.82	1.41	2	0.49
Station 3	-0.49	0.10	0.24	68.82	0.23	2	0.89
Station 1	-0.41	0.11	0.26	68.82	0.89	2	0.64
Station 5	-0.22	0.13	0.42	68.82	0.07	2	0.96
Station 6	0.04	0.11	0.10	68.82	0.49	2	0.78
Station 7	0.23	0.10	0.63	68.82	1.49	2	0.47
Station 2	0.24	0.10	-0.16	68.82	2.37	2	0.31
Station 4	0.34	0.11	-0.14	68.82	4.25	2	0.12
Station 8	0.42	0.09	-0.82	68.82	2.99	2	0.22
Station 11	0.47	0.09	0.35	68.82	1.76	2	0.41
Station 10	0.74	0.08	0.65	68.82	1.11	2	0.57

Table 2: Individual stations fit (by item difficulty order)

Test of fit for individual items are also illustrated using graphical representation. This is particularly crucial for a data set with a large number of students where chi square statistics tend to be overly sensitive in detecting misfit. Figure 2 shows an example of a graphical representation for station scores with excellent fit to the model, station three Critical Care and O&G. Note that the observed mean scores received by students in the high, medium and low ability groups were plotted and compared with the expected pattern of scores curve by the model.

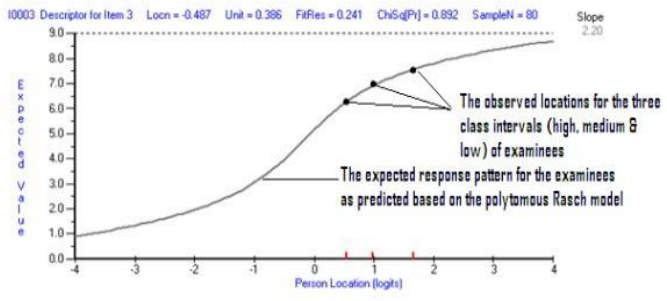

Figure 2: Item with excellent fit: Station 3 critical care and O&G

Measure of reliability

Rasch analysis provides a measure of reliability in the form of PSI, which is similar in concept to the Cronbach's Alpha under CTT and the G coefficient under Generalisability Theory. PSI is simply the ratio of estimated true variance for the measure to the total observed variance, calculated using linear interval scores (the logit scores) which exclude the extreme raw scores. This is in contrast to the Cronbach's Alpha and G coefficient computed using non-linear raw scores with extreme scores included and hence contains an infinitely large standard error.[10]

PSI for the OSCE data in this study is 0.704 indicating that 70.4% of the variance in the observed scores for students is due to the estimated true variance in students' levels of clinical competence. This figure also indicates that the error variance which includes examiner severity is 29.6%.

Measures of individual station difficulty

Location estimates in Column 2, Table 2 indicates the estimated task difficulty for each station. Clinical tasks for station ten appear to be the most difficult, and tasks in station nine are the easiest. The metric calibration of task difficulty based on the PRM is sample *independent.* In other words, they are derived by comparing task difficulty and student ability on a common continuum, that is the underlying construct. These estimates of station difficulty are therefore criterion-referenced. These are useful meta-data to be included in the OSCE item bank to facilitate continuity and equity in the OSCE exam standard.

Measures for students' clinical competency

When the OSCE data fits the Rasch model, the RUMM 2020 programme transforms ordinal raw scores into a metric linear interval scale using the unit of logits, as shown in Column 3, Table 3. This process is commonly called Rasch scaling. In addition to providing measures for individual students' clinical competence, Rasch scaling also provides direct estimates of standard error for each estimate of student clinical competence (Column 4). These individualised standard errors provide quantification for the precision of every individual's measure. They can be used to describe the range within which each student's true clinical competence may be located. As compared to the application of an average standard error for all students'

scores, the direct estimate of standard errors provided in Rasch analysis is a more justifiable and accurate way of establishing precision of measurement. This is particularly valuable in making decisions to award a particular grade or in determining progression for students whose scores fall within the borderline area. After all, the reality in all assessment data is that standard error of measurement varies across the range of student ability.[11]

ID	Total	Locn	SE	Residual	DF
12	60	0.17	0.22	-0.5	9.5
74	61	0.219	0.22	-1.658	9.5
57	63	0.324	0.23	-0.644	9.5
42	63	0.324	0.23	0.207	9.5
5	63	0.324	0.23	-0.526	9.5
72	63	0.324	0.23	-0.01	9.5
37	64	0.38	0.24	0.029	9.5
58	65	0.438	0.24	-0.286	9.5
64	65	0.438	0.24	-0.054	9.5
21	66	0.499	0.25	0.303	9.5
52	66	0.499	0.25	-0.802	9.5
2	66	0.499	0.25	0.172	9.5
31	67	0.562	0.25	0.239	9.5
7	67	0.562	0.25	#-3.283	9.5
.....					
22	83	1.904	0.32	0.175	9.5
75	83	1.904	0.32	-0.959	9.5
43	84	2.009	0.33	-0.357	9.5

#: fit residual value exceeds limit set for test-of-fit

Table 3: Individual person fit and location estimates (excerpt) – by location order

Identification of anomalies in individual student scores

The unit of analysis in Rasch analysis is 'individual' student scores. This is in contrast to CTT where the unit of analysis is 'group' scores. As a result of this feature, Rasch analysis is able to flag anomalies in the score patterns across the stations for individual students. Fit residual statistics are used to flag individual student scores misfit to the model, as shown in Column 5 of the Rasch analysis output in Table 3. The general rule of thumb in the interpretation of fit residual statistics is that a fit residual value beyond the range of + / - 2.5 indicates some anomalies in the individual students' scores pattern. This is an important piece of information for the course coordinator, clinical teacher and/or the OSCE station developer. These fit residual statistics serve as a starting point for further investigation as to the reasons behind the anomalies which could be due to data entry error, examiner's bias, individual student's physical conditions such as fatigue, sickness, test anxiety etc. With these insights, appropriate actions can then be undertaken to account for and possibly resolve any issues that might be relevant.

Discussion

As demonstrated in the preceding result sections, Rasch modelling seems a practical quality assurance tool for OSCE data.

Rasch scaling provides measures of students' performance that are criterion-referenced to the clinical task assessed. Therefore the results are generalisable and meaningful to guide learning and instructions. Individual station difficulty estimates are also criterion-referenced and not sample-dependent. These can be included as metadata for all stations in the OSCE item bank. With this criterion-referenced data on individual OSCE station difficulty, different OSCE stations across medical disciplines and OSCE stations targeting different level of training can be linked effortlessly through the co-calibration of test items and test linking and/or test equating.

In light of the establishment of an OSCE item bank with meta-data from Rasch analysis, standard setting for OSCEs will become less cognitively demanding for judges as compared to standard setting methods based on CTT.[12] As a result, standard setting for OSCEs will also become a more time-efficient process.[12, 13]

Ultimately, medical schools can aspire for the integration of Rasch modelling in scaling and psychometric evaluation for both written assessments (MCQs, EMQs, and Short Answer Questions) and performance assessments such as OSCE, Mini-CEX, Professional Portfolio, Clinical Audit etc. This is the path towards establishing one common scale (one ruler), to link all different test forms and formats, i.e. horizontal tests linking and equating. A similar scale can

also be used to link assessment data across different stages of training for vertical test linking and equating towards a competency-based curriculum and assessment framework.

Conclusion

With retrospective analysis of an OSCE data set, we have illustrated that Rasch modelling based on the PRM provides a formal test of unidimensionality of the underlying construct across multiple stations in clinical examinations such as the OSCE. We have also exemplified how Rasch analysis establishes evidence for construct validity of OSCE sum-scores. Also discussed in the preceding sections are long term benefits of the application of Rasch analysis for OSCE data, in particular, and other components of summative assessment in general.

References

1. Lunz ME, Stahl JA. Impact of examiners on candidate scores: An introduction to the use of multifacet Rasch model analysis for oral examinations. Teach Learn Med. 1993; 5(3): 174-181.
2. Linacre JM. A user's guide to Facets Rasch measurement computer program, version 3.66.0. 2009. Chicago: Winsteps.com.
3. Bhakta B, Tennant A, Horton M, Lawton G, Andrich D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. BMC Med Educ (serial on the Internet). 2005; 5(9); [Cited 28 March 2011]. Available from: http://www.biomedcentral.com/content/pdf/1472-6920-5-9.pdf
4. Andrich D. Rasch models for measurement. Beverly Hills, CA: Sage Publication; 1988.
5. Marais I, Andrich D. Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. J Appl Meas; 2008, 9(2): 105–124.
6. Andrich D, Sheridan B, Lyne A, Luo G. RUMM: a windows-based item analysis program employing Rasch unidimensional measurement models. Perth: Murdoch University; 2004.
7. Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43: 561-573.
8. Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press; 1982.
9. Wright BD, Stone MH. Best test design: Rasch Measurement. Chicago: MESA Press; 1979.
10. Schumacker RE, Smith EV. A Rasch perspective. Educ Psychol Meas. 2007; 67(3): 394-409.
11. Qualls-Payne A. A comparison of score level estimates of the standard error of measurement. Journal of Educational Measurement. 1992; 29: 213-225.
12. Wang N. Use of the Rasch IRT model in standard setting: an item-mapping method. Journal of Educational Measurement; 2003, 40(3): 231–253.
13. MacCann RG, Stanley G. The use of Rasch modelling to improve standard setting. Practical Assessment, Research & Evaluation. 2006; 11(2), [Cited 15 January 2010]. Available from http://pareonline.net/pdf/v11n2.pdf.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Michele Gawlinski's contributions to the early stages of this project.

PEER REVIEW

Not commissioned. Externally peer reviewed

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

FUNDING

None.

ETHICS COMMITTEE APPROVAL

This project has been approved by the Human Research Ethics Committee, the University of Notre Dame Australia.