British Journal of Medicine & Medical Research
17(12): 1-7, 2016, Article no.BJMMR.29295
ISSN: 2231-0614, NLM ID: 101570965

SCIENCEDOMAIN *international*
www.sciencedomain.org

Scoring Short Answer Questions of Five Borderline Medical Students

Monique Trigg[1*], John Barnard[1], Hannah Pham[2] and Peter Devitt[2]

[1]Excel Psychological and Educational Consultancy (EPEC), P.O.Box 3147, Doncaster East, Victoria, 3109, Australia.
[2]Department of Surgery, University of Adelaide, Adelaide, South Australia, 5005, Australia.

Authors' contributions

This work was carried out in collaboration between all authors. Authors HP and PD compiled the questions and marked the answers. Authors MT and HP administered the tests and authors JB and MT analysed the data. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJMMR/2016/29295
Editor(s):
(1) Syed Faisal Zaidi, Department of Basic Medical Sciences, College of Medicine, King Saud Bin Abdulaziz University-HS, National Guard Health Affairs, King Abdulaziz Medical City, Kingdom of Saudi Arabia.
Reviewers:
(1) Abidullah Khan, Khyber Teaching Hospital, Peshawar, Pakistan.
(2) Naina Kumar, MMIMSR, Mullana, Ambala, Haryana, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/16310

Short Research Article

Received 1st September 2016
Accepted 19th September 2016
Published 23rd September 2016

ABSTRACT

Background: The assessment of medical knowledge is integral to becoming a medical practitioner in Australia, and Short Answer Questions (SAQs) are frequently used in this process. This paper compares the use of Classical Test Theory (CTT) and the Rasch rating scale measurement framework in scoring SAQs to evaluate the competence of borderline candidates in Australian medical students.
Aims: The aim of this study was to utilise two scoring paradigms to compare the results of borderline medical students on SAQs.
Methods: Forty SAQs were administered to 140, fifth year medical students at an Australian university in an online practice examination. Aligned with CTT, each student's performance was expressed as the sum of the question scores. The data was then also analysed within the Rasch rating scale measurement framework and measures of performance were obtained. The two sets of results were compared across borderline students.
Results: According to CTT, five students were identified as being exactly at the pass mark of 50 per cent. Rasch analysis indicated however that although the students had the same ability

**Corresponding author: E-mail: Monique@EPECat.com;*
Co-author: E-mail: John@epecat.com;

estimates, their approach to answering SAQs were vastly different, altering the interpretation of their overall performance.

Conclusion: The sole use of CTT in the analysis of examination data may result in issues of validity and reliability when measuring clinical competence. The Rasch rating scale measurement framework may be invaluable in informing the analysis of performance in high stakes scenarios to ensure fair decisions of clinical competence.

Keywords: Assessment; Rasch measurement; short answer question; classical test theory; medical education.

1. BACKGROUND

A standard method used to obtain an overall score of a test or examination is to simply add the scores on the questions. This is based on a framework known as Classical Test Theory (CTT) [1]. Such methods however, disregard the subjective nature of the data by making unwarranted assumptions [2]. One assumption is that the data is on an interval scale (i.e. that the relative value of each response category across questions are the same and the unit increases across the rating scale are equal in value). In other words, it is assumed that each question contributes just as much to the total score as any other question and that all questions are equally difficult [3]. It is further assumed in CTT that the scores within a question are equally spaced, i.e. that the difference between an increase from a score of '1' to '2' is the same as the difference between an increase from a score of '2' to '3'.

Rasch measurement theory provides an alternative approach to the analysis of data [4,5]. The family of Rasch models are based on the idea that data must conform to some reasonable hierarchy of 'less than/more than' on a single continuum of interest [6]. The Rasch model uses the traditional total score as a starting point for estimating probabilities of responding. The model is based on the simple idea that all persons are more likely to answer easy items correctly than difficult items, and all items are more likely to be passed by persons of high ability than those of low ability [7].

The Rasch model provides estimates for each question (difficulty) and each person (ability) separately, but on the same scale; something that is not possible in Classical Test Theory [3,6]. Equality of intervals is achieved through log transformations of raw data odds, and abstraction is accomplished through probabilistic equations [8]. The person ability and question difficulty estimates, having been subjected to a log transformation, are displayed along a *logit* (log odds unit) scale which is an interval scale in which the unit intervals have a consistent value or meaning.

When questions are scored according to a marking guide, an extension of the basic Rasch model is needed to accurately represent such polytomous data [6]. Rating scale analysis allows each question's relative difficulty to be estimated, as well as the pattern of the scale categories in each question to yield a rating scale structure. Thus, each question has a difficulty estimate, and the scale itself also has a series of thresholds [9].

As stated above, one of the main advantages of Rasch measurement theory over classical (traditional) theory is that item difficulty estimates and person ability estimates can be located on a common interval level scale [3]. This can also be done for rating scales, where the difficulty to 'achieve' each category can be shown on a scale [10].

2. METHODS

The study was designed to compare the performance of Australian medical students across two scoring regimes, namely, Classical Test Theory (CTT), and the Rasch rating scale measurement framework.

The study was conducted at an Australian University on 12 September 2015 after approval by the Human Research Ethics Committee. The sample consisted of 140, fifth Year MBBS students (85 females and 55 males). Participation was voluntary, and students were informed about the aims and the nature of the research prior to registering their interest to participate online via an application on the Google cloud platform, *Google Forms*. Students gave written consent to participate when registering.

A unique username and password was generated for each participant in order to access

the practice examination. These were only provided to participants on the day of their examination, upon sign-in. The examination consisted of 40 Short Answer Questions (SAQs), of which 20 were marked on a four-point scale (0 to 3); ten questions were marked on a three-point scale (0 to 2) and ten questions were marked on a two-point scale (0 or 1) to yield a maximum possible score of 90. The questions were administered through an application on the Google cloud platform, *Google Forms* and included the eight disciplines; medicine (med), surgery (surg), psychiatry (psych), orthopaedics (ortho), general practice (GP), obstetrics / gynaecology (O&G), paediatrics (paed) and anaesthetics / pain medicine / intensive care (APIC).

Students were allowed 90 minutes to complete the test and all responses were captured online and collated for scoring and analysis. The responses were distributed to three independent markers, which were marked according to a marking guide.

3. OVERALL RESULTS

3.1 Classical Test Theory

The average of the three markers' scores was calculated for each question and rounded to the nearest whole number. Each student's performance was expressed as the sum of the question scores. Table 1 summarises descriptive statistics of students' total scores. The results show a minimum score of 23.3 per cent, a maximum score of 76.7 per cent and a mean of 52.2 per cent.

Table 1. Descriptive statistics

	N	Min.	Max.	Mean	Std. dev.
Total	140	21.0	69.0	46.96	9.9

If 45 out of 90 is considered as the passing score (i.e. 50%), then there were five students exactly at the passing score. These five students' scoring patterns are considered in more detail below. The Cronbach alpha reliability of 0.77 yields a standard error of measurement of 4.7 which can be used to determine a precision band around the passing score, and this average standard error is applied in the same way for all students' scores.

3.2 Rasch Rating Scale

In the Rasch rating scale analysis, the first step was to determine whether the questions were well targeted to the students. Fig. 1 shows a good match between the student ability measures (red bars) and the 40 question difficulties (blue bars).

The mapping confirms good targeting with the questions providing maximum information around the peak of the student ability estimates. Since each question also had a number of score categories (two to four), more detailed information about the targeting was obtained by plotting the individual category difficulties of each question against the student ability estimates; see Fig. 2.

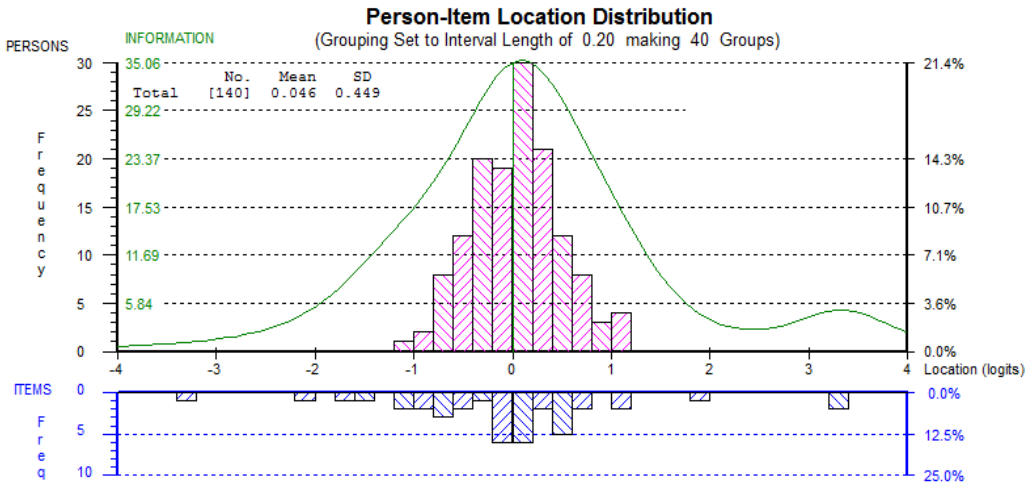

Fig. 1. Person-item location distribution

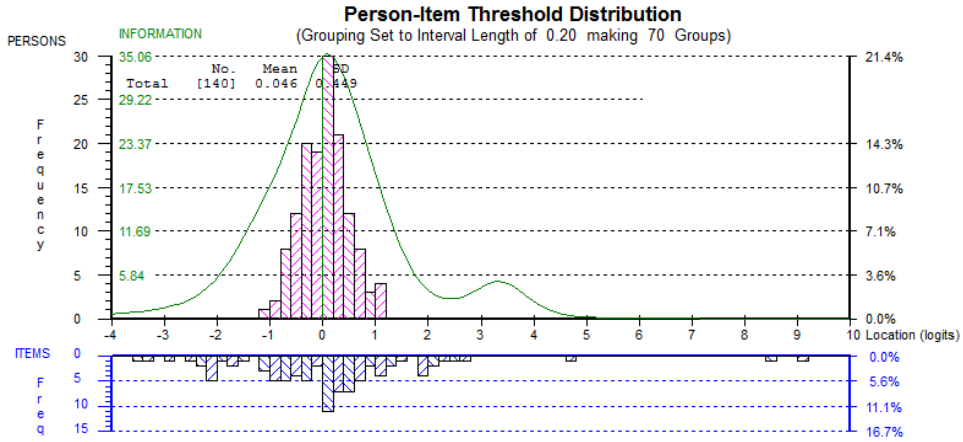

Fig. 2. Person-item threshold distribution

Good targeting was confirmed and it is noted that there were three (especially two) categories that were very difficult to achieve, i.e. to obtain such scores.

The power of test-of-fit was investigated next, and the Person Separation Index (PSI) of 0.77 indicated that 77 per cent of the variance in the observed scores was due to the estimated true variance in students' levels of clinical competence and that the error variance, which includes marker severity, is 23 per cent. The mean student ability of 0.046 logits (Standard deviation of 0.449) matched the mean question difficulty of 0.000 logits (standard deviation of 1.210) very well. The student mean fit residual of 0.019 (SD of 0.853) and the question mean fit residual of 0.290 (SD of 0.792) showed slightly more misfit in the question estimates and the chi-square probability value of the question-trait interaction indicated that Rasch analyses could be done.

The analysis confirmed significant differences between the overall question difficulties, ranging from -3.251 logits (question 10) to 3.381 (question 35). It was the most difficult to score 3 in question 35 (category difficulty of 5.762 logits) followed by a three in question 14 (category difficulty of 5.197 logits) whilst it was the easiest to score a one in question 27 (category difficulty of -2.036 logits).

The step structures of the questions were used to explore the scoring structures in more detail. Question 2, for example, suggested that scoring in all categories was not always probable (see Fig. 3). The scoring structure showed that it was never most likely to score 1 (red line) in this question on a scale of 0 to 3. Students with lower ability estimates most likely scored zero (blue line) after which a score of 2 (green line) and then 3 (purple line) became more likely as ability estimates increased.

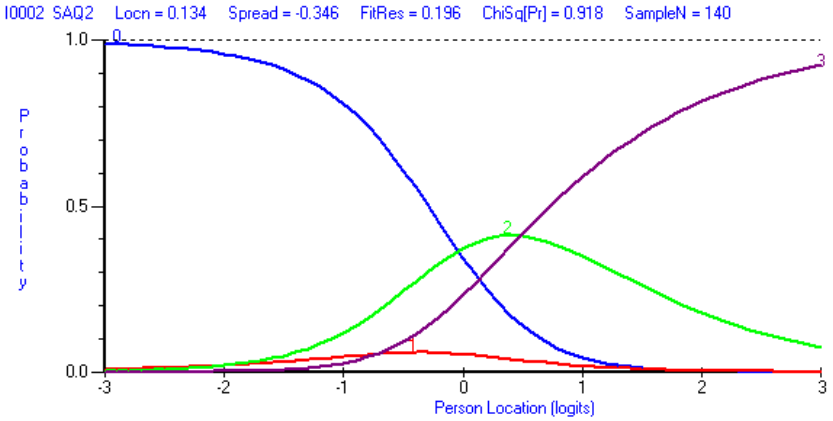

Fig. 3. Scoring structure of question 2

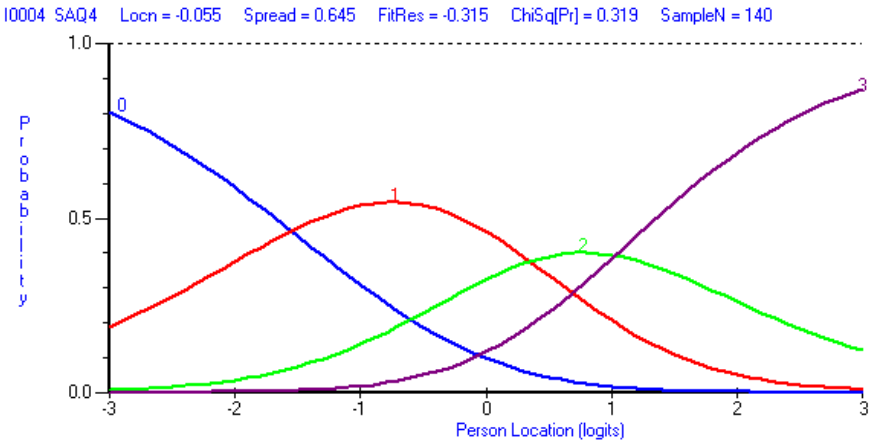

Fig. 4. Scoring structure of question 4

In contrast, some questions such as question 4 were much better ordered as can be seen in above Fig. 4.

3.3 Results of Borderline Students

According to CTT, five students (N1 to N5) were identified as being exactly at the pass mark of 50 per cent, with a score of 45 out of 90. The distribution of their marks over the eight disciplines can be seen in more detail in Table 2.

In the Rasch analysis a score of 45 equates to an ability estimate of -0.047 logits with a SEM of 0.207.

Table 2. Borderline student scores by discipline

Discipline	N1**	N2	N3	N4	N5
Med (21*)	11	9	6	7	9
Surg (17)	9	11	11	11	7
Paed (11)	7	5	9	8	6
O&G (11)	6	6	1	2	6
GP (9)	1	5	6	6	6
Psych (9)	4	3	4	4	5
APIC (9)	5	4	5	5	5
Ortho (3)	2	2	3	2	1

**: Maximum possible score in discipline*
***: Student 1*

As seen in the table above, although the students had the same total scores of 45 out of 90, their performance was very different when discipline scores were examined in detail. Student 1 (N1) had a score of 11 per cent in General Practice (GP), while the four other borderline students had a score of at least 56 per cent as can be seen in Table 2. If GP was considered a fundamental area of knowledge in determining the clinical competence of a student, it can be argued that this student should have failed. If 50% was the cut-score for each of the eight disciplines, student 1 would have passed six disciplines, student 2 four and the other students five each. If a pass in at least six disciplines was added as a criterion to pass, only student 1 would have passed.

Although the conclusions above provide additional information about the performance of the students, the assumption is that the difficulty over the disciplines is constant, i.e. that 50% in GP is the same as 50% in Med. Furthermore, it is noted that the number of questions over disciplines is not the same. Requiring 50% to pass in GP is thus different to requiring 50% to pass in Ortho.

When the data fits the Rasch model, ordinal raw scores are converted into a metric linear interval scale using the unit of logits. Such measures can subsequently be used to compare performances directly because all measures are expressed on the same scale. Calibration of all questions in the test yielded a mean question difficulty of 0.00 logits (by definition) with a standard deviation of 1.210 (as mentioned above). The mean student ability was 0.046 logits with a standard deviation of 0.449. From a score equivalence table it was derived that 50% overall equates to an ability estimate of -0.047 logits.

The exam was subdivided into the eight disciplines and the mean difficulty of each discipline was calculated. These are summarised

in the table below. It is noted that there is quite significant differences in the mean discipline difficulties. Surgery was the "most difficult" and Medicine the "easiest"; a difference of 0.88 logits which clearly indicates that a score of 50% in one discipline is not the same as a score of 50% in another.

Table 3. Mean difficulty of the eight disciplines in logits

Med	-0.30	GP	0.36
Surg	0.58	Psych	0.15
Paed	0.17	APIC	0.20
O&G	-0.06	Ortho	-0.06

A first step in investigating a student's performance is to consider the actual responses to the questions. The overall response patterns can be expressed in terms of a fit residual to obtain an indication of the extent that they are aberrant. The fit residuals are shown for each student in the table below.

Table 4. Individual person fit estimates

Student	Fit residual
N1	0.34
N2	1.12
N3	-0.14
N4	-0.89
N5	-0.09

It is noted that the biggest difference is between students 2 and 4 and therefore these two students will be further considered in more detail. In the following table the performance of student 2 is compared with the performance of student 4 in logits for each discipline.

Table 5. Performance of students 2 and 4 in logits by discipline

Discipline	N2	N4
Med	-0.22	-0.58
Surg	0.92	0.92
Paed	-0.36	0.91
O&G	0.37	-1.57
GP	0.28	0.64
Psych	-0.99	-0.43
APIC	-0.49	0.01
Ortho	0.97	0.97

Student 2 passed four of the eight disciplines with a mean of 0.06 logits whilst student 4 passed five disciplines with a mean of 0.11 logits.

4. DISCUSSION

Passing or failing borderline candidates has been and perhaps, will always be a contentious issue. This study demonstrates that simply adding scores on individual questions to obtain an overall score may be misleading, especially if subsets of scores are used to make pass/fail decisions. Although sum scores give some indication of relative performance, subset scores are not on a common scale and should therefore not be compared directly. Scores on different scales are likely to lead to biased interpretation whereas the location of measures on a single scale, as common practice in Rasch measurement, overcomes such potential bias. Through exploring the scoring structures of the questions and Rasch calibration to construct a single scale, valid comparisons can be made. The Rasch rating scale measurement framework provides a myriad of benefits in analysing data in high stakes examination situations due to the ability to provide more detailed information on individual performance.

Corroborating the findings of Tor and Steketee[4], the use of Rasch modelling in assessing clinical competence in medical students can provide much needed quality assurance in high stakes examinations.

The certification process in medical education often requires candidates to pass multiple forms of assessment. Through the application of Rasch measurement theory in the psychometric analysis of these, it is possible to create one common scale to locate performance across all assessments. This may allow regulating bodies and verification authorities to maintain that, the same standard is required to pass any form of a certification exam, at any point in time.

Although Rasch analysis is not sample-dependent, it is noted that the sample in this study was from one Australian University. It is not envisaged however, that other samples in the target population of medical graduates would differ greatly to the current sample. In addition, the examination was devised as a practice formal examination only, with question content developed by content experts. In this cause, the range of 2.01 in the fit residual statistics may be accounted for since students knew it was not a critical examination.

5. CONCLUSION

A comparison of CTT and Rasch analysis on the SAQ data of borderline medical students evidenced that Rasch provides long term advantages in assessment in medical education through providing critical information on individual score patterns and the assessment of clinical competence on SAQs. Future research could utilise the Rasch model to examine differences in individual ability estimates when individuals are given the choice to select which SAQs / Cases they respond to.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Royal K, Gilliland K, Kernick E. Using Rasch measurement to score, evaluate, and improve examinations in an anatomy course. Anat Sci Educ [EJ1044124]. 2014; 7(6):450-460.
2. Magno C. Demonstrating the difference between classical test theory and item response theory using derived test data. Int J Educ Psychol Assess [ED506058]. 2009;1(1):1-11.
3. Downing S. Item response theory: Applications of modern test theory in medical education. Med Educ [PMID: 12945568]. 2003;37(8):739-745.
4. Tor E, Steketee C. Rasch analysis on OSCE data: An illustrative example. AMJ. 2011;4(6):339-345. Available:http//dx.doi.org/10.4066/AMJ.2011.755
5. McNamara T, Knoch U. The Rasch wars: The emergence of Rasch measurement in language testing. Lamguage Testing. 2012;29(4):553-574.
6. Barnard JJ. A primer on measurement theory. Melbourne: Excel Psychological and Educational Consultancy; 2012.
7. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed). New Jersey: Lawrence Erlbaum Associates.
8. Linacre M. Winsteps tutorial. (Retrieved November 7, 2014) Available:http://www.winsteps.com/tutorials.htm
9. San Martin E, del Pino G, de Boeck P. IRT models for ability-based guessing. Appl Psychol Meas. 2006;30(3):183-203.
10. Andrich D. Rating formulation for ordered response categories. Psychometrika. 1978;43(4):561-573.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/16310