Available online at www.sciencedirect.com

ScienceDirect

Journal of the Chinese Medical Association 74 (2011) 125−129

www.jcma-online.com

Original Article

Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model

Shun-Chin Yang [a,b,†], Mei-Yung Tsou [a,b,†], En-Tzu Chen [a,b], Kwok-Hon Chan [a,b], Kuang-Yi Chang [a,b,c,*]

[a] *Department of Anesthesiology, Taipei Veterans General Hospital, Taipei, Taiwan, ROC*
[b] *National Yang-Ming University School of Medicine, Taipei, Taiwan, ROC*
[c] *Division of Biostatistics, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan, ROC*

Received May 27, 2010; accepted August 4, 2010

Abstract

Background: Test evaluation in a clinical curriculum is important for medical education. To identify examinee ability and appropriateness of the test content, this study used the Rasch model to analyze an examination in anesthesiology for medical students.
Methods: Fifty items were administered to 119 fifth- and sixth-year medical students in the exam. The Rasch model was used to perform item analysis of the examination. Misfit items or examinees were excluded first, then test reliability was assessed with reliability indices. Both examinee ability measures and item difficulty were estimated and expressed in a common logit unit, which could be further translated into probability of correct responses in the examination.
Results: After the exclusion of two misfit items and one misfit examinee, the estimated test reliability was only 0.63. The mean item difficulty was set at 0 by definition (SD = 2.02) and the mean examinee ability was 1.56 (SD = 0.71), which means that the examinees were able to correctly answer 83% of items on average. There were 21 items with difficulty lower than the least able examinee and two items with difficulty higher than the most able one.
Conclusion: We demonstrated that statistical item analysis with the Rasch model could provide valuable information related to test reliability, item difficulty and examinee ability, which could be applied to further item modification and future test development of clinical curriculums for medical students.

Keywords: Anesthesiology; Item response theory; Multiple choice question; Rasch model; Reliability

1. Introduction

Evaluation of learning and teaching in a clinical curriculum with examinations is important for medical education. It can assess the effect of a teaching program and the levels of clinical knowledge absorbed by medical students. An examination should precisely and reliably measure proficiency of students and be able to discern examinees with different levels of ability. In order to assure the validity and reliability of an examination, items in an examination should be subject to thorough investigation with some psychometric methods. Although item analyses are common in many tests, it was unusual for examinations in anesthesiology for medical students.

Item response theory (IRT) has been extensively applied to miscellaneous types of test analyses.[1−3] The theoretic foundation and mathematical characteristics of IRT overcome some limitations in classical test theory, and IRT has gained

[*] Corresponding author. Dr. Kuang-Yi Chang, Department of Anesthesiology, Taipei Veterans General Hospital, 201, Section 2, Shih-Pai Road, Taipei 112, Taiwan, ROC.
E-mail address: kychang@vghtpe.gov.tw (K.-Y. Chang).
[†] Shun-Chin Yang and Mei-Yung Tsou contributed equally to this work.

1726-4901/$ - see front matter Copyright © 2011 Elsevier Taiwan LLC and the Chinese Medical Association. All rights reserved.
doi:10.1016/j.jcma.2011.01.027

popularity in a variety of disciplines.[1,3] Among various models related to IRT, the one-parameter Rasch model possesses the properties of fundamental measurement and is commonly used to develop and validate various instruments.[1,4] Applications of the Rasch model in medicine are also increasing rapidly.[5−7] The Rasch analysis compares examinee ability and item difficulty by fitting them on the same continuum through calibrating processes. An advantage of the Rasch model over other competing models is that relatively few subjects are required to obtain useful estimates with reasonable precision.[8] Therefore, we conducted this study to perform item analysis of an examination in anesthesiology for medical students in Taiwan with the Rasch analysis. Through the analysis, item difficulty and examinee ability could be identified in the same scale and reliability of the examination and fitness of items could be assessed. The results can be used as a reference for formulating questions and constructing item banks of similar exams in the future.

2. Methods

2.1. Data

The data were taken retrospectively from the final results of an exam in anesthesiology for medical students in Taiwan in 2008. The analytic protocol was approved by the institutional review board (VGHIRB No. 97-08-14A). All of the fifth- or sixth-year medical students who took anesthesiology as a compulsory subject had to take the examination at the end of the course. The analysis was based on the responses of 119 candidates to 50 items on the written examination. All items were multiple choice questions with five options and single best answer. Examinees had to complete the exam in 1 hour, and all of them accomplished the objective in time. If an item was answered correctly, the examinee would get two points. No punishment was given for a wrong answer except no point gotten. The number of correctly answered items multiplied by two for an examinee would equal his original score, and the response of an examinee to an item was recoded into 1 or 0 based on right or wrong answer, respectively. After recoding the original responses into binary data, the transformed data were submitted to the Rasch analyses.

2.2. Statistical analysis

2.2.1. Introduction to the Rasch model

The Rasch model is the simplest form among the IRT models. It is a logistic model of probability for monotonically increasing functions. It presents the simple relationship among examinee ability (θ), item difficulty (b), and the probability of a correct response (p) with the following formula:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i - b_j,$$

where p_{ij} = the probability of examinee i passing item j; θ_i = the ability of the ith examinee; b_j = the difficulty of the jth item.

The left side of the formula involves the log transformation of the odds of correctly answering a particular item. For examinees with very low ability, the probability of “passing” an item is virtually 0. As examinee ability approaches the difficulty of the item, the probability of a correct response increases gradually. When examinee ability matches the item difficulty, the probability of a correct response is 0.5. Finally, examinees with very high ability have virtually a 1.0 probability of a correct response. The ability measures and item difficulties can be expressed in the same logit unit. The difference between the examinee ability and item difficulty can be translated directly into the probability of correct response. For example, an examinee with ability of 1 logit unit higher than the difficulty of an item will be expected to answer the item correctly with a probability of 0.73.

2.2.2. Exclusion of misfit items and examinees

Fit statistics of items and examinees were checked at first and misfit items or examinees should be excluded from further analyses due to the violation of model assumption or redundancy.[9] Two types of fit statistics were provided in this study: the mean square (MSQ) and standardized fit statistics (ZSTD). Both fit statistics can be calculated in two versions: variance-weighted and unweighted.[10] For an item, weighted fit statistics are more important and the acceptable range of weighted MSQ is from 0.8 to 1.2,[10,11] and ZSTD values are between −2 and 2. For an examinee, unweighted fit statistics is of interest and the value of unweighted ZSTD should not be greater than 5.[12]

2.2.3. Evaluation of test reliability

After exclusion of misfit items and examinees, test reliability was evaluated to ensure the consistency of the estimated results. Two reliability coefficients were provided in the analysis—the reliability and separation indices. The reliability index provided in Rasch analysis is conceptually analogous to the Cronbach alpha.[1] Separation index is equal to the square root of reliability divided by (1—reliability), which represents how well the exam can distinguish examinees in terms of their ability location.[13] A separation index of 1.5 is considered an acceptable level of separation capacity for a test and indices of 2.0 and 3.0 indicate good and excellent levels of separation capacity, respectively.[14,15] If the test reliability was unsatisfactory, the effect of increasing items on reliability was also assessed with the Spearman−Brown prophecy formula.[16]

2.2.4. Estimation of examinee ability and item difficulty

The examinee ability and item difficulty could be converted into linear interval measures with the logit (log odds) transformation by the Rasch model. The mean item difficulty was assigned the logit value of 0 because the difference between examinee ability and item difficulty is relative rather than absolute. Examinee ability was then estimated in relation to item difficulty. Estimates with higher values in the logit scale suggest higher examinee ability or item difficulty. An item distribution map was constructed to illustrate the distribution of the examinee ability and item difficulty on the same scale.

The original scores and logit scales from the Rasch analysis of examinees are presented as mean with standard deviation (SD). The Rasch analyses were performed with Winsteps software, Version 3.68 (Winsteps.com, Chicago, IL, USA). Other analyses were conducted with SPSS version 15.0 (SPSS Inc., Chicago, IL, USA).

3. Results

The mean original score of the examinees was 71.3, with SD of 8. Figure 1 illustrates the histogram of original scores. Obviously, the distribution of original scores did not deviate from normality. Figure 2 shows two kinds of item fit statistics of the initial Rasch analysis. According to the aforementioned criteria, misfit items were excluded one by one until no misfit item was identified, and then misfit examinees were eliminated from further analysis. After fit statistic analysis, one examinee and two items (items 20 and 45) were excluded. Weighted MSQ fit indices of the remaining items lay between 0.87 and 1.11, and unweighted MSQ values of the remaining examinees were in the range from 0.26 to 3.92.

After the exclusion of misfit examinee and items, the estimated test reliability and separation indices were 0.63 and 1.3, respectively. Figure 3 depicts the hypothetic test reliability by increasing item numbers with the Spearman−Brown formula. Given the test reliability of 0.63, at least 70 items were necessary for a test reliability of 0.7, and doubling the exam length by adding items with the same properties would give a test reliability of 0.77. To achieve a test reliability of 0.8, at least 120 items were required in the examination.

The mean item difficulty was set at 0 by definition and the standard deviation of item difficulty was 2.02. The mean examinee ability was 1.56 with SD of 0.71. Figure 4 presents the results of item analysis. Among the remaining 48 items, only item 4 was answered correctly by all the examinees. The item difficulties lay between −3.46 and 5.13, and the range of examinee ability was from −0.6 to 3.43. Figure 2 maps the distribution of examinee ability and item difficulty on the same logit scale. The distribution of examinee ability did not

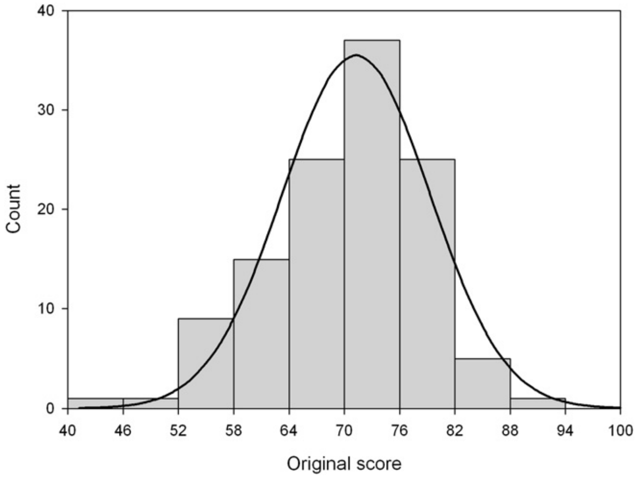


Fig. 1. Histogram of original scores in the examination in anesthesiology for medical students.

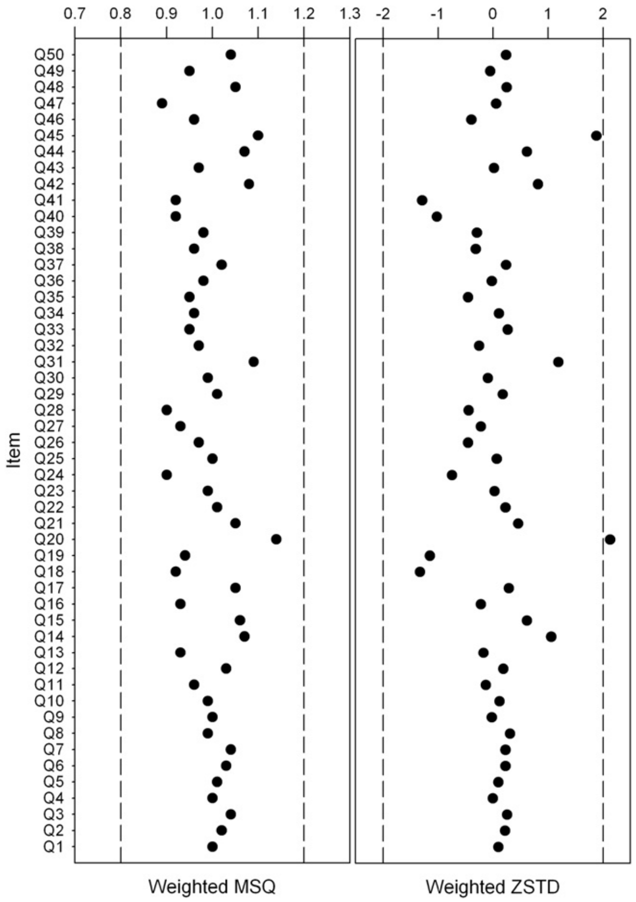


Fig. 2. Weighted MSQ and ZSTD of each item in the Rasch analysis. This figure illustrates the result of initial item fit statistic analysis. One item (Item 20) has a fit statistic value out of the range of pre-specified criteria for weighted ZSTD. Therefore, this item would be excluded from the second round of item fit statistic analysis. Similarly, another misfit item (item 45) was excluded in the second round of fit statistic analysis. The processes would be continued until no misfit item was identified. After the exclusion of all misfit items, misfit examinee would also be excluded in a similar manner. MSQ = mean square; ZSTD = standardized fit statistics.

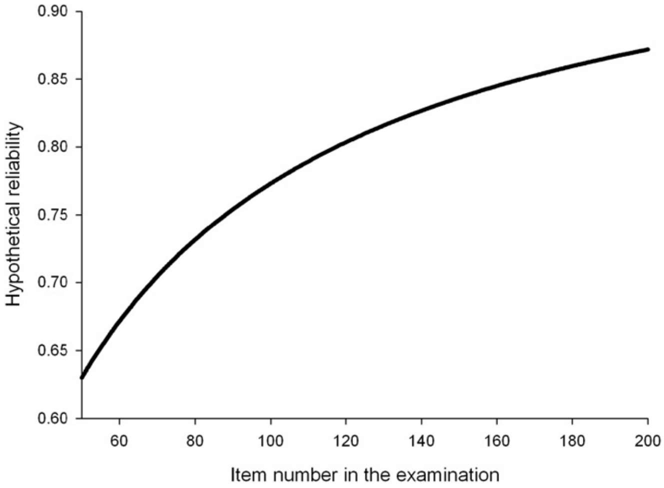
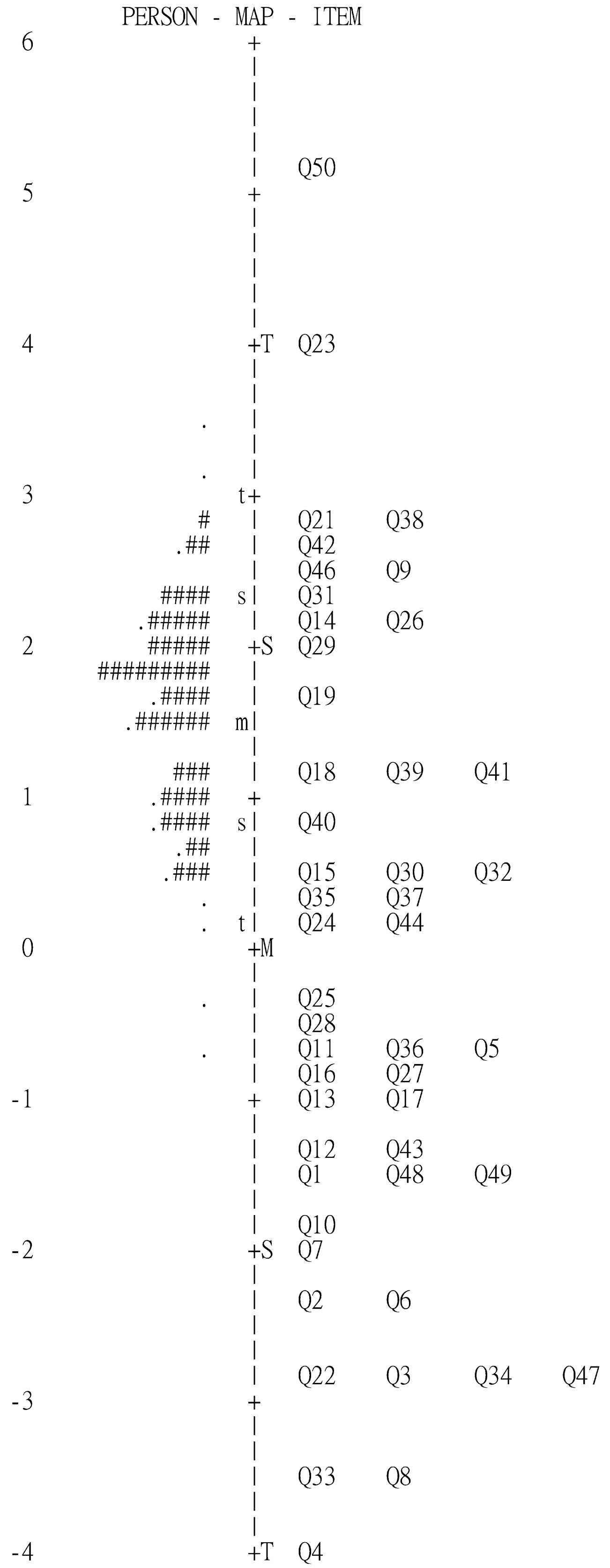


Fig. 3. Effects of increasing item numbers on the hypothetic test reliability estimated by Spearman−Brown prophecy formula.



```
          PERSON - MAP - ITEM
 6                     +
                       |
                       |
                       |
                       |
                       |  Q50
 5                     +
                       |
                       |
                       |
                       |
                       |
 4                     +T Q23
                       |
                       |
                    .  |
                       |
                    .  |
 3                    t+
                    #  |  Q21    Q38
                  .##  |  Q42
                       |  Q46    Q9
                 #### s|  Q31
               .#####  |  Q14    Q26
 2              #####  +S Q29
           ##########  |
                .####  |  Q19
              .###### m|
                       |
                  ###  |  Q18    Q39    Q41
 1              .####  +
                .#### s|  Q40
                  .##  |
                 .###  |  Q15    Q30    Q32
                    .  |  Q35    Q37
                    . t|  Q24    Q44
 0                     +M
                       |
                    .  |  Q25
                       |  Q28
                    .  |  Q11    Q36    Q5
                       |  Q16    Q27
-1                     +  Q13    Q17
                       |
                       |  Q12    Q43
                       |  Q1     Q48    Q49
                       |
                       |  Q10
-2                     +S Q7
                       |
                       |  Q2     Q6
                       |
                       |
                       |  Q22    Q3     Q34    Q47
-3                     +
                       |
                       |
                       |  Q33    Q8
                       |
                       |
-4                     +T Q4
```

Fig. 4. Item distribution map for examinee and item measures on the same scale. Each "#" represents two examinees and "." represents one examinee. The vertical dash line and the leftmost figures represent the common logit scale of examinee ability and item difficulty. The uppercase letters "M", "S" and "T" on the right side of the dash line represents the "mean", "one standard deviation" and "two standard deviations" of item difficulty estimates, respectively. The lowercase letters "m", "s" and "t" on the left side of the dash line represents the "mean", "one standard deviation" and "two standard deviations" of examinee ability estimates, respectively. The figures with the letter "Q" at the right side of the vertical dash line represent item numbers in the exam.

obviously deviate from normality. In contrast, the distribution of item difficulty distinctly diverged from normality. There were 21 items with difficulties lower than the ability of least able students and only two items with difficulties higher than the ability of most proficient student. The remaining 25 items possessed difficulty within the range of examinee ability distribution. Although nearly 62% of students had ability measures ranged between 0.8–2.1, there were only five items with difficulty in this range.

4. Discussion

After item analysis with the Rasch model, we had several findings which provided valuable information for improvement of the examination. First, there were only two misfit items, which indicated most of the items did not violate assumptions of the Rasch model,[9] which indicated acceptable item development quality because most items behaved as anticipated—students with higher ability measures could be expected to have higher probability of having correct answers to these items. Second, the test reliability was an unsatisfactory 0.63, which means that the test results were not so reliable. To improve the test reliability, increasing the item numbers should be considered. Third, the examination was relatively easy for most of the students. To enhance the discrimination of the test, item difficulty should be adjusted to promote usefulness of the exam. These implications could be used as a reference or guide for future test development or item bank construction of similar exams. For example, the test developer of easier items should be informed to modify the content of items to upgrade the difficulty. For misfit items, their developer should be notified to investigate the sources of misfit, like controversy over the correct answers or ambiguity in the item descriptions.

Tests can be classified into two major groups: norm-referenced (relative) and criterion-referenced (absolute) tests.[17,18] The goal of a norm-referenced test is to classify students. In contrast, a criterion-referenced test aims to judge how well examinees are doing relative to a pre-determined performance level (criterion). In fact, this examination was a criterion-referenced test. To conduct item analysis, we had to use a norm approach because almost all test analytic methods are norm-based. Although the current trend for examinations in clinical curriculums is criterion-referenced evaluation, a concomitant norm-referenced assessment, such as the Rasch analysis, can provide valuable and objective information related to test reliability, item difficulty and examinee ability. Results of an examination contain more useful information than whether an examinee should pass or not. Through our analytic processes, such information could be extracted for future test development in clinical curriculums.

Generally, items with difficulty equal to the range of examinee ability would provide most information for parameter estimation.[19,20] Among the remaining 48 items, twenty-one items had difficulty below the ability of least able examinee. Moreover, items with comparable difficulties were not uncommon in this exam. According to Schumacker,[9] supernumerary items with similar difficulty and variance provided little additional information for parameter estimation. Besides, as many as 62% of examinees had ability measures between 0.8–2.1. However, there were only five items with difficulty in this range. It would be difficult to differentiate students with ability in this range due to lack of comparable items. The inadequacy of item difficulty which resulted in less useful information for parameter estimation at least partially accounted for the low test reliability. Further efforts could be exerted to improve the test discrimination and reliability. For example, increasing the item numbers in the test or modifying the difficulty of items may be useful, as we demonstrated in the analysis.

Quality control is important for test development. Performance in an examination should reflect only the proficiency in the aimed construct, not other irrelevant ones, which means that the test should be unidimensional.[21] This is also a basic assumption of the Rasch model. Fit statistics can be used to evaluate unidimensionality and as a measure to evaluate the validity of examinees' responses.[12,22] Among all the items, only two items with fit statistics exceeded the pre-determined range of fitness. It indicated that most of the developed items fitted the Rasch model well, which implied passable item development quality and acceptable construct validity of the exam.

There are some limitations in our study. First, this study was a cross-sectional survey. The findings in this study represent a single period. To evaluate trend and variation in examinations in anesthesiology over time, longitudinal follow-up data should be collected and analyzed. Second, the Rasch model is the simplest form in item response theory. Several alternative models could also be considered for conducting item analysis. The analysis of fit statistics confirmed that our data fitted the Rasch assumptions well and it was reasonable to perform item analysis with the Rasch model. Since the Rasch model could achieve our analytic goals, applying other more sophisticated models to item analysis was beyond the scope of this study.

In conclusion, we demonstrated that the Rasch analysis could be applied satisfactorily to item analysis of the examination in anesthesiology for medical students. It also provided valuable information for further item modification and future test development in clinical curriculums.

Acknowledgment

This study was supported by grants from Taipei Veterans General Hospital (V97A-109) and Anesthesiology Research and Development Foundation, Taipei, Taiwan (ARDF9905).

References

1. Bond TG, Fox CM. *Applying the Rasch model: fundamental measurement in the human sciences*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum Associates; 2007, pp. 29–66.
2. Embretson SE, Reise SP. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000, pp. 273–305.
3. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications; 1991, pp. 7–28.
4. Andrich D. *Rasch models for measurement*. Newbury Park, CA: Sage Publications; 1988, pp. 16–23.
5. Decruynaere C, Thonnard JL, Plaghki L. How many response levels do children distinguish on faces scales for pain assessment? *Eur J Pain* 2009; **13**:641–8.
6. Hawthorne G, Densley K, Pallant JF, Mortimer D, Segal L. Deriving utility scores from the SF-36 health instrument using Rasch analysis. *Qual Life Res* 2008;**17**:1183–93.
7. Lamoureux EL, Pesudovs K, Pallant JF, Rees G, Hassell JB, Caudle LE, et al. An evaluation of the 10-item vision core measure 1 (VCM1) scale (the core module of the vision-related quality of life scale) using Rasch analysis. *Ophthalmic Epidemiol* 2008;**15**:224–33.
8. Linacre JM. Sample size and item calibration stability. *Rasch Meas Trans* 1994;**7**:328.
9. Schumacker RE. Rasch measurement: the dichotomous model. In: Smith EV, Smith RM, editors. *Introduction to Rasch measurement: theory, models and applications*. Maple Grove, MN: JAM Press; 2004, pp. 226–53.
10. Sheu C, Chen C, Su Y, Wang W. Using SAS PROC NLMIXED to fit item response theory models. *Behav Res Methods* 2005;**37**:202–18.
11. Wright B, Linacre J, Gustafson J. Reasonable mean-square fit values. *Rasch Meas Trans* 1994;**8**:370.
12. Wright BD, Stone MH. *Best test design: Rasch measurement*. Chicago: MESA Press; 1979, pp. 165–9.
13. De Ayala R. *The theory and practice of item response theory*. New York: Guilford Press; 2008, pp. 407–8.
14. Fisher WP. Reliability statistics. *Rasch Meas Trans* 1992;**6**:238.
15. Wright BD, Master GN. *Rating scale analysis*. Chicago: MESA Press; 1982, pp. 90–117.
16. Allen MJ, Yen WM. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing; 1979, pp. 72–91.
17. Gregory RJ. *Psychological testing: history, principles, and applications*. 5th ed. Boston: Pearson/Allyn and Bacon; 2007, pp. 76–95.
18. Murphy KR, Davidshofer CO. *Psychological testing: Principles and applications*. 6th ed. Upper Saddle River, NJ: Pearson/Prentice Hall; 2005, pp. 436–7.
19. Hambleton RK, Swaminathan H. *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing; 1985, pp. 101–23.
20. Lord FM. *Applications of item response theory to practical testing problems*. Hillsdale, IN: Lawrence Erlbaum Associates; 1980, pp. 65–77.
21. Bonnel AM, Boureau F. Labor pain assessment: validity of a behavioral index. *Pain* 1985;**22**:81–90.
22. Smith Jr EV. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *J Appl Meas* 2001;**2**:281–311.