British Journal of Anaesthesia **104 (6):** 717–22 (2010)
doi:10.1093/bja/aeq097 Advance Access publication April 28, 2010

Item analysis for the written test of Taiwanese board certification examination in anaesthesiology using the Rasch model

K.-Y. Chang[1 2], M.-Y. Tsou[1 2], K.-H. Chan[1 2], S.-H. Chang[2], J. J. Tai[2] and H.-H. Chen[2]*

[1]Department of Anaesthesiology, Taipei Veterans General Hospital and School of Medicine, National Yang-Ming University, Taipei, Taiwan, Republic of China. [2]Division of Biostatistics/Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Room 533, No. 17, Hsuchow Road, Taipei, Taiwan 100, Republic of China

**Corresponding author. E-mail: chenlin@ntu.edu.tw*

Background. On the written test of board certification examination for anaesthesiology, the probability of a question being answered correctly is subject to two main factors, item difficulty and examinee ability. Thus, item analysis can provide insight into the appropriateness of a particular test, given the ability of examinees.

Methods. Study subjects were 36 Taiwanese examinees tested with 100 questions related to anaesthesiology. We used the Rasch model to perform item analysis of questions answered by each examinee to assess the effects of question difficulty and examinee ability using a common logit scale. Additionally, we evaluated test reliability and virtual failure rates under different criteria.

Results. The mean examinee ability was higher than the mean item difficulty in this written test by 1.28 (SD=0.57) logit units, which means that the examinees, on average, were able to correctly answer 78% of items. The difficulty of items decreased from 4.25 to −2.43 on the logit scale, corresponding to the probability of having a correct answer from 5% to 98%. There were 60 items with difficulty lower than the least able examinee and seven difficult items beyond the most able one. The agreement of item difficulty between test developers and our Rasch model was poor (weighted κ=0.23).

Conclusions. We demonstrated how to assess the construct validity and reliability of the written examination in order to provide useful information for future board certification examinations.
The study was approved by the institutional review board with the following trial registered number: VGHIRB No. 97-08-14A.

Br J Anaesth 2010; **104**: 717–22

Keywords: anaesthesiology; certification examination; Rasch model; reliability

Accepted for publication: March 10, 2010

The purpose of the board certification examination in anaesthesiology is to evaluate whether an examinee is able to exceed minimum requirement for clinical practice. More importantly, test items must be able to measure examinee ability with a high degree of precision and accuracy, and discriminate good performance from bad. An examination where all items were extremely difficult, or conversely, extremely simple, is clearly undesirable. Thus, how to evaluate the difficulty of a particular question, given examinee ability, is a high priority. As both examinee ability and item difficulty are abstract constructs, how to calibrate these two latent variables on the same scale plays a crucial role in analyses of the written test for a certifying examination in anaesthesiology. One of the approaches is to consider the Rasch model, which has been used to analyse a variety of standardized examinations[1–4] and validate various instruments in clinical practice.[5–7] To the best of our knowledge, it has not been applied to board certification examination in anaesthesiology yet.[8–10] We applied the Rasch model to data taken from the written test of board certification examination for anaesthesiologists in order to quantify examinee ability and item difficulty and also to evaluate test reliability. We then assessed the agreement of item difficulties as rated by the test developers and results obtained from the Rasch analysis. We also simulated different scenarios where numbers of very difficult and easy items defined by logit score derived from the Rash analysis were deleted to investigate their influences on test reliability with respect to item difficulty and examinee ability.

Methods

Study subjects and data collection

The data were taken from the final results of the September 2007 board certification examination for anaesthesiology in Taiwan. Eligibility criteria for the examination included completion of an approved residency programme at an accredited medical training centre, and demonstration of comparable experience in clinical practice. The board certification examination for anaesthesiology in Taiwan consists of two stages, where candidates are required to pass the first stage (written test) to become eligible for the second stage (oral examination). The analyses described below were based on the responses of 36 candidates to 100 items on the written examination and their performance on the oral examination. The items in the written and oral examinations were developed and reviewed by a committee of eight anaesthesiology professionals. The item categories and number of items in each category are presented in Appendix 1. Each test developer provided 12 items for written examination and one item for oral examination, based on the assigned categories. All developed items were reviewed thoroughly by all committee members, with six items in the written examination being deleted after the review. Ten additional items were selected from the latest item bank of board certification examination to maintain the number of total questions at 100. All items in the written examination were multiple-choice questions with five answer options and single best answer.

All examinees were required to complete the written test within 2 h, which all did successfully. For each correctly answered item, a value of one was assigned; no penalties were given for incorrect answers, which were assigned a value of zero. The number of correctly answered items represented the original total score.

After passing the written examination, examinees had to undertake the oral examination for completing a full course of the board certification examination in anaesthesiology. However, as our interest in this study focused on the assessment of item difficulty in the written test making allowance for examinee ability, the detailed analysis on oral examination was not performed.

Statistical analyses

We first used the Rasch model to assess the effects of person ability and item difficulty on the probability of a correct response for that item using the following equation:[11 12]

$$P_{ij}=\frac{\mathrm{e}^{\left(\theta_i-b_j\right)}}{1+\mathrm{e}^{\left(\theta_i-b_j\right)}}$$

where P_{ij} is the probability of the person i answering the jth item correctly; θ_i the logit score for ability of the ith person; and b_j the logit score for difficulty of the jth item. In such a way, the ability measures and item difficulties can be expressed using the same logit units on a common scale. The difference between examinee ability and item difficulty is directly related to the probability of a correct response for that item. For example, if the ability of an examinee is higher than the difficulty of a specific item by 1 logit unit, the probability of answering correctly is 0.73 [$\mathrm{e}^{(1)}/(1+\mathrm{e}^{(1)})$]. When the examinee ability is comparable with the item difficulty, the difference between two logit scores is 0, which means that he or she has 50% probability of having a correct or wrong answer. Since the contrast between examinee ability and item difficulty is relative rather than absolute, it is customary to set the mean item difficulty in logit unit as 0. The details of statistical technique on the Rasch model refer to the previous statistical literature.[11 12]

Misfit items that did not meet the standard criteria of fit statistics were excluded from further analyses ($0.8<$ weighted mean square<1.2;[13 14] $-2<$standardized fit statistics<2).[15] The person and item reliability indices were calculated to ensure consistency using two forms of reliability coefficients, reliability (analogous to Cronbach's α) with the value between 0 and 1 and separation index (the number of statistically different performance strata that the test can identify in the sample),[16] being 1.50, 2.00, and 3.00 for three levels of separation: acceptable, good, and excellent, respectively.[17] Although the person and item reliability indices were used to denote the replicability of person ranking for a parallel test and item location for another sample of examinees with comparable ability, respectively,[18] the person reliability is of major concern for a certifying examination.

An item distribution map was constructed to illustrate the distribution of persons and items on the same logit scale (Fig. 1). A virtual failure rate of the written examination could be estimated by setting different criteria based on examinee ability in logit unit. Examinee ability lower than the specified criteria was doomed to failure in the virtual analysis. Since items with difficulty within the range of examinee ability were the most informative,[19] we assessed whether item reduction based on the results of the Rasch analysis could provide a consistent estimation of examinee ability by comparing Pearson's correlation coefficients under three conditions, condition I excludes only misfit items, condition II eliminates very easy and difficult items, and condition III deletes all items out of the ability range of examinees.