Available online at www.sciencedirect.com

ScienceDirect

Journal of the Chinese Medical Association 77 (2014) 150–154

www.jcma-online.com

Original Article

Comparison of proficiency in an anesthesiology course across distinct medical student cohorts: Psychometric approaches to test equating

Shu-Wei Liao, Kuang-Yi Chang, Chien-Kun Ting, Mei-Yung Tsou, En-Tzu Chen, Kwok-Hon Chan, Wen-Kuei Chang*

Department of Anesthesiology, Taipei Veterans General Hospital and National Yang-Ming University School of Medicine, Taipei, Taiwan, ROC

Received July 29, 2013; accepted August 27, 2013

Abstract

Background: Examinations are necessary for assessment of student proficiency in medical education, but comparison of achievement across different cohorts in different tests is challenging. We applied psychometric test equating methods to compare student proficiency in two different examinations for a clinical anesthesiology course.

Methods: Each examination contained 50 multiple choice items and nine common items were identified from the two examinations (administered in 2011 and 2012). The common item design was used for test equating. Two psychometric test-equating approaches, chained linear equating and item response theory, were used to compare student proficiency in anesthesiology across distinct medical student cohorts. Raw scores from the 2012 test were linearly transformed to the 2011 scale using the chained method, and then Rasch analysis was applied to calibrate examinee ability and item difficulty in the two examinations on a common scale.

Results: Both the linear equating method and Rasch analysis indicated that students in the 2011 examination performed better than those who took the 2012 examination (both $p < 0.001$). Rasch analysis revealed that the range of student ability was between −0.53 and 4.16, while the difficulty of all items ranged from −5.25 to 6.32. No significant difference in mean item difficulty was noted among the common items and other items in the two examinations.

Conclusion: Although both the chained linear equating method and Rasch analysis can be readily applied to practical test-equating issues in medical education, Rasch analysis exhibited more versatility in test parameter estimation and item bank development for clinical curriculums.

Keywords: anesthesiology; linear equating; multiple choice question; Rasch model; test equating

1. Introduction

Examinations have great value in medical education for assessing the performance of medical student on a clinical curriculum.[1,2] However, finding the most effective method for comparing achievement between different cohorts of students is a challenge for practitioners. In general, performance on a clinical test is expressed in terms of raw scores, and direct comparison of student ability across distinct cohorts is not feasible when different tests are administered.[3,4] For example, a student with lower ability may obtain a higher raw score on an easier test than another student with higher ability who takes a more difficult examination. This makes direct comparison of raw scores unreasonable and unreliable. Although traditional equating methodologies developed on the basis of classic test theory can be used to overcome this issue, the assumption that examinees are from identical population is disputable in various cases.[5,6]

The development of item response theory (IRT) provides a promising solution to test equating,[7] and IRT models have been successfully applied to test equating problems for health status measures[8] and medical licensing examinations.[9] One IRT model, the one-parameter Rasch model, can be readily

Conflicts of interest: The authors declare that there are no conflicts of interest related to the subject matter or materials discussed in this article.

* Corresponding author. Dr. Wen-Kuei Chang, Department of Anesthesiology, Taipei Veterans General Hospital, 201, Section 2, Shih-Pai Road, Taipei 112, Taiwan, ROC.

E-mail address: wkchang@vghtpe.gov.tw (W.-K. Chang).

1726-4901/$ - see front matter
http://dx.doi.org/10.1016/j.jcma.2013.10.011

applied in comparing examinee performance not only in the same test but also in different tests, provided there is sufficient linkage between the tests. This is because Rasch analysis has great potential and flexibility for the management of systematic missing data.[10] In this study, we applied both the traditional linear equating method and Rasch analysis for practical comparison of student performance in different examinations of a clinical curriculum using a common item design. The versatility of Rasch analysis can also be demonstrated by additional contributions to parameter estimation of item difficulty and examinee ability, and to further item bank development.

2. Methods

2.1. Data collection

The data were collected in 2011 and 2012 from written examinations in anesthesiology for students in their fifth year of medical education in a university in Taiwan. The examinations were administered after the end of a compulsory course in anesthesiology to assess the understanding of basic concepts of the curriculum. A portion of the data collected in 2011 were used in a previous study.[2] There were 130 and 122 students taking the examinations in 2011 and 2012, respectively. Both examinations comprised 50 multiple choice questions (MCQs) with five options and a single best answer. The time limit in each case was 1 hour, and all examinees completed the task within the allotted time. After completing the test, examinees were asked to return their examination paper and corresponding answer sheet to ensure test question security. Two marks were given for each correct response, and no penalty was assigned for a wrong answer. Accordingly, the number of items correctly answered by an examinee multiplied by two was equal to his or her raw score, and these item response data were used for test linking and equating.

2.2. Statistical analysis

Among the 100 items in the two examinations, nine common items were identified and thus a common item design with concurrent calibration was applied for test equating. The chained linear equating method was used to transform raw scores in the 2012 examination to the 2011 scale.[11] Examinee ability and item difficulty were calibrated using Rasch analysis, and estimated examinee and item parameters were expressed using the logit (log odds) unit on a common linear scale.[1] More details on the chained linear equating method and Rasch model are available in Appendix 1 and the literature.[4,5] The raw scores, linearly transformed scores, and logit scores obtained from the linear equating process and Rasch analysis are presented as mean ± SD. The linear relationship between transformed and raw scores in the 2011 examination was plotted and an item distribution map was constructed to illustrate the examinee ability distribution and item difficulty in the two examinations on the common scale. Rasch analysis was also used to estimate test reliability. Comparisons of parametric data between the two student groups were conducted using an independent t test or one-way analysis of variance as appropriate. Responses to common items between the two student groups using a χ^2 test and differential item functioning analysis using logistic regression were also compared to preclude the possibility of test security compromise. Winsteps software, Version 3.75 (Winsteps.com, Chicago, IL, USA) was used for Rasch analyses. The default setting for mean item difficulty was 0 on the common logit scale to avoid scale indeterminacy. Larger estimated values indicate higher examinee ability and item difficulty. Differential item functioning and other analyses were performed using PASW Statistics version 18.0 (SPSS Inc., Chicago, IL, USA). A p value less than 0.05 was considered statistically significant.

3. Results

The common test reliability of the two examinations was 0.77. Table 1 compares the mean original score, number of correct responses to common items, linearly transformed scores on the 2011 scale, and Rasch logit scores between students in the 2011 and 2012 examinations. All comparisons reveal that students in the 2011 test performed significantly better than those in the 2012 examination (all $p < 0.001$). Fig. 1 illustrates the linear relationship between transformed and raw scores in the 2012 examination. Note that transformed scores on the 2011 scale were lower than the corresponding raw scores in the 2012 examination. Rasch analysis further revealed that the mean item difficulty for the common and other items in the 2011 and 2012 examinations was −0.53, 0, and −0.46, respectively. No significant difference in mean item difficulty was noted among the common items and the remaining items in the 2011 and 2012 examinations ($p = 0.59$).

Fig. 2 presents the common item distribution map according to Rasch analysis. The mean ability of all examinees in logit units was 1.94 ± 0.99. The range of examinee ability was distributed over the interval between −0.53 and 4.16. The difficulty of all items ranged from −5.25 to 6.32 with SD of 2.17. Only three items exceeded the examinee ability range and 41 items had difficulty lower than the ability of the least able examinee.

Table 2 compares the proportion of correct responses to common items between students in the 2011 and 2012

Table 1
Comparison of raw scores, number of correct response to common items, linearly transformed 2012 score, and Rasch logit scores between the two student groups.

	2011 (n = 130)	2012 (n = 122)	p
Raw score	81.2 ± 6.5	73.3 ± 7.4	<0.001
Correct responses to common items	7.4 ± 0.7	6.2 ± 0.9	<0.001
Linearly transformed score	–	70.6 ± 8.2	<0.001
Rasch logit score	2.64 ± 0.72	1.19 ± 0.61	<0.001

Data are presented as mean ± SD.

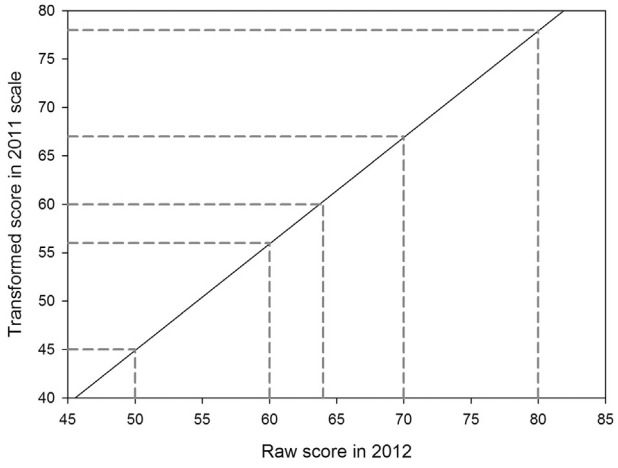

Fig. 1. Linear relationship between raw scores in the 2012 examination and scores transformed to the 2011 scale. Raw scores in the 2012 examination were linearly transformed to the 2011 examination scale using the chained linear equating method. For example, raw scores of 50, 60, 70, and 80 in the 2012 test correspond to linearly transformed scores of 45, 56, 67, and 78, respectively, on the 2011 scale. If the pass threshold was set at a raw score of 60 in the 2011 examination, it should be increased slightly to a raw score of 64 in the 2012 examination on the same benchmark scale.

examinations and assesses potential item differentiation to prevent compromise of test security. Among the nine common items, only three items show a significant difference in the proportion of correct responses. Note that students in the 2012 examination performed better than their 2011 counterparts on only one item (Q40R39). However, further item differentiation analysis revealed that the 2012 student group had no advantage over those who took the 2011 test ($p = 0.19$). One differential item (Q40R43) was observed (see Appendix 2 for detailed information).

4. Discussion

We applied the classical linear equating method and Rasch analysis for test linking and equating in medical education. Our investigation has provided several important findings. First, both methods can be readily used to compare student performance in different tests with minimal assumptions. Note that on average, students in the 2011 examination performed better than their 2012 counterparts. Since the two examinations were of comparable difficulty, the underlying causes of discrepancy in performance should be thoroughly investigated. Our approaches not only compare the achievements of distinct student groups but also provide clues for possible educational gaps in different student cohorts, which merits further study to ensure the educational quality of clinical curriculums.[9] Second, even when only a relatively small sample size is available, Rasch analysis can still provide efficient and reliable estimation.[10,12] Although more complicated models can also be considered in test equating, a larger sample size is needed to generate stable estimates. However, such a requirement is rarely met in test equating processes conducted in a single medical university. Third, the analytical

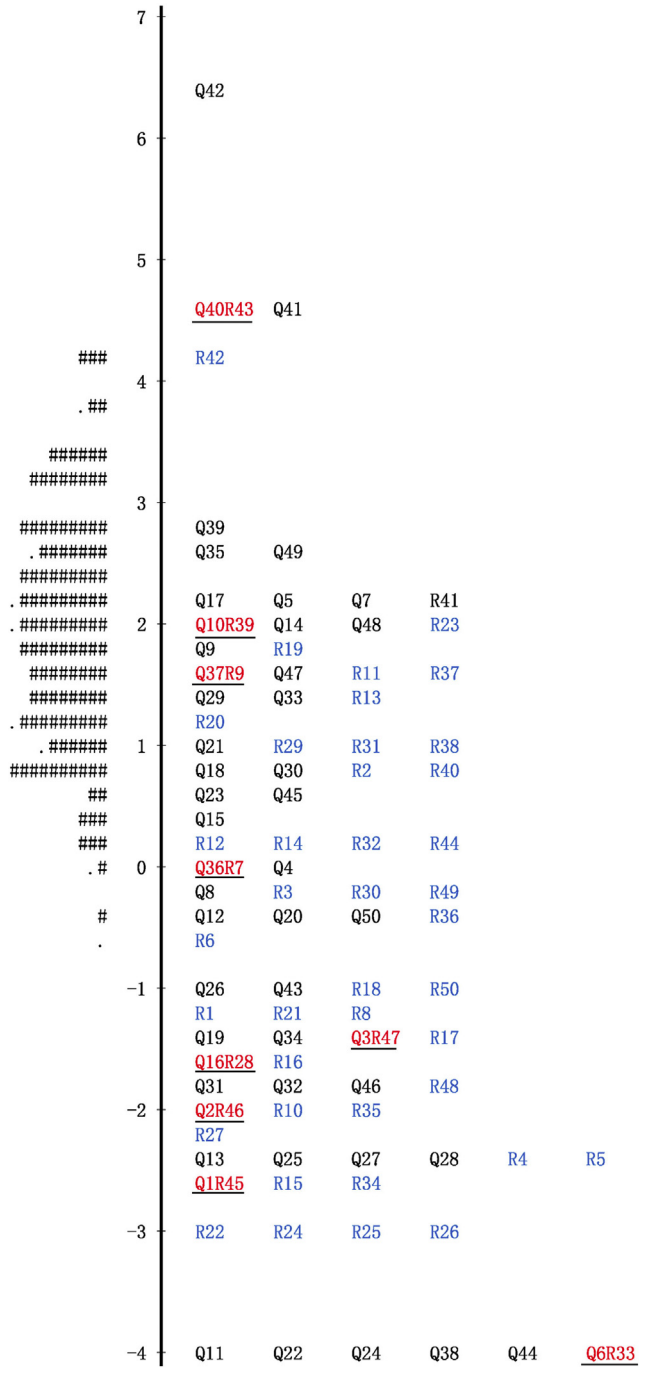

Fig. 2. Item distribution map for the examinations held in 2011 and 2012. Q and R are prefixes for item numbers in the 2011 and 2012 examinations, respectively. Items in the 2011 and 2012 examinations are in black and gray font, respectively. Common items are underlined. The mean item difficulty was set to 0 by default. The scale on the vertical line represents the common logit unit for item difficulty and examinee ability according to Rasch analysis. The distribution of examinee ability is illustrated on the left-hand side of the common scale using # to denote two students and • to denote one student.

results can be used to develop an item bank for anesthesiology in clinical education owing to the availability of item parameter estimates. Test equating processes can combine the item responses for two or more examinations and yield estimates of item parameters on a common scale to construct an

Table 2
Comparison of the proportion of correct responses to common items between the two examinations held in different years.

Common item	Correct responses (%)		p	Item differentiation p
	2011	2012		
Q1R45	100.0	96.7	0.054	0.999
Q2R46	98.5	95.9	0.269	0.880
Q3R47	95.4	95.1	0.910	0.150
Q6R33	100.0	100.0	–	–
Q10R39	43.1	55.7	0.045	0.185
Q16R28	97.7	94.3	0.205	0.782
Q36R7	100.0	64.8	<0.001	0.999
Q37R9	97.7	13.9	<0.001	0.153
Q40R43	10.0	8.2	0.619	<0.001

Of the three common items that significantly differ in the proportion of correct responses, examinees in 2012 performed better than their counterparts in 2011 for only one item (Q10R39). However, item differentiation was not observed for this item.

item bank for more advanced applications such as computerized adaptive testing.[13]

There are three common types of IRT-based test equating method: separate calibration, calibration with fixed common item parameters, and concurrent calibration.[14,15] Separate calibration estimates item parameters for different test forms separately and then computes transformation coefficients using parameter estimates for the common items. For calibration with fixed common item parameters, item parameters are estimated in one of the test forms and then the parameters are fixed for common items to estimate item parameters in the other test forms to achieve a common scale. In concurrent calibration, all item and examinee parameters in multiple test forms are simultaneously estimated in a single run. As suggested by Hanson and Béguin,[15] concurrent calibration generally results in fewer errors than the other methods and thus we adopted this approach in our study.

The development of item banks is a time-consuming and laborious task. In practice, the number of items in a test is finite and item parameters estimated in separate tests cannot be directly used to construct an item bank since the parameters are estimated on different bases.[16] As a result, test linking and equating are essential for the development of an item bank. After an item bank is constructed, it is necessary to maintain and improve item bank quality through regular updating by adding new items and removing those that are out of date. Accordingly, a highly efficient test equating method is indispensable for sustainable use of an item bank. Our approach provides a feasible solution to practical test equating issues encountered in the development and maintenance of item banks for clinical curriculums.

There are some limitations in the study. First, we only used chained linear equating and the Rasch model for test equating. Although other psychometric methods can also be considered, the linear equating method is technically easy to conduct and the Rasch model was a better choice to provide stable and reliable results in our study setting.[2] Second, the test equating process was conducted in a single university and the generalization of the item parameters and item bank obtained needs further test equating processes to link up test forms developed in other schools. Nevertheless, our study provides feasible approaches for test equating in clinical curriculums.

In conclusion, we demonstrated the applicability of chained linear equating and Rasch analysis to practical test equating issues for clinical curriculums. Student performance in different tests can be compared using these methodologies through the linkage of common items. The item parameter estimates can also be applied to the development of an item bank for a clinical curriculum. Rasch analysis provides a versatile solution to test equating and linking problems in medical education.

Acknowledgments

This study was supported by grants from Taipei Veterans General Hospital (V102B-020) and the Anesthesiology Research and Development Foundation, Taipei, Taiwan (ARDF10003).

Appendix 1. Brief introduction to the chained linear equating method and the Rasch model

Chained linear equating

Linear equating methodology can be applied to test equating through the linkage of common items in two tests. Chained linear equating transforms a raw score in the new form X into a rescaled score in the base form Y, given the assumption of population invariance, according to

$$y = \mu(\mathrm{Y}) + \frac{\mathrm{SD_Y SD_{XC}}}{\mathrm{SD_X SD_{YC}}}[x - \mu(\mathrm{X})] + \frac{\mathrm{SD_Y}}{\mathrm{SD_{YC}}}[\mu_{\mathrm{X}}(\mathrm{C}) - \mu_{\mathrm{Y}}(\mathrm{C})]$$

where $\mathrm{SD_x}$, $\mathrm{SD_Y}$, $\mathrm{SD_{xC}}$, and SD_{YC} are the standard deviation for the raw scores of forms X and Y and for raw scores calculated from common items in forms X and form Y, respectively, and $\mu(\mathrm{X})$, $\mu(\mathrm{Y})$, $\mu_{\mathrm{x}}(\mathrm{C})$, and $\mu_{\mathrm{Y}}(\mathrm{C})$ are mean raw scores for forms X and Y and mean raw scores calculated from common items in forms X and Y, respectively.

Rasch model

The Rasch model translates the probability of a correct item response into person ability and item difficulty according to

$$P\left(Y_{ij} = 1 \middle| \theta_i, b_j\right) = \frac{\exp\left(\theta_i - b_j\right)}{1 + \exp\left(\theta_i - b_j\right)}$$

where $P(Y_{ij} = 1)$ is the probability that examinee i will correct answer item j, θ_i is the ability of examinee i, and b_j is the difficulty of item j. Rasch analysis can estimate examinee ability and item difficulty parameters on the same scale and the difference between examinee ability and item difficulty can be directly expressed as the probability of a correct response.

Because the common item Q40R43 shows a differential item function, further evaluation of the response pattern is necessary.

Appendix 2.
Comparison of Rasch logit scores between correct and incorrect groups for item Q40R43 stratified by examination year

	Incorrect		Correct		*p*
	n	Logit score	*n*	Logit score	
2011	117	2.54 ± 0.66	13	3.52 ± 0.66	<0.001
2012	112	1.21 ± 0.60	10	1.00 ± 0.65	0.31

Comparison of the Rasch logit score (mean ± SD) reveals a significant difference between correct and incorrect respondents. This finding suggests that the item functioned normally in the 2011 examination because correct respondents obtained higher scores. By contrast, in the 2012 examination, no significant difference in logit score was noted between the correct and incorrect subgroups. This implies that item Q40R43 cannot discriminate proficient from unskilled students.

References

1. Yang SC, Tsou MY, Chen ET, Chan KH, Chang KY. Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model. *J Chin Med Assoc* 2011;**74**:125−9.
2. Huang YF, Tsou MY, Chen ET, Chan KH, Chang KY. Item response analysis on an examination in anesthesiology for medical students in Taiwan: a comparison of one- and two-parameter logistic models. *J Chin Med Assoc* 2013;**76**:344−9.
3. Osterlind SJ. *Modern measurement: theory, principles, and applications of mental appraisal*. Boston, MA: Allyn & Bacon/Pearson; 2010.
4. De Ayala RJ. *The theory and practice of item response theory*. New York: Guilford Press; 2009.
5. Kolen MJ, Brennan RL. *Test equating, scaling, and linking: methods and practices*. New York: Springer; 2004.
6. Holland PW, Dorans NJ. Linking and equating. In: Brennan RL, editor. *Educational measurement*. 4th ed. Westport, CT: Greenwood; 2006. p. 187−220.
7. Muraki E, Hombo CM, Lee YW. Equating and linking of performance assessments. *Appl Psychol Meas* 2000;**24**:325−37.
8. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000;**38**:II43−59.
9. Yim MK, Huh S. Test equating of the medical licensing examination in 2003 and 2004 based on the item response theory. *J Educ Eval Health Prof* 2006;**3**:2.
10. Chang KY, Tsou MY, Chan KH, Chen HH. Application of the Rasch model to develop a simplified version of a multiattribute utility measurement on attitude toward labor epidural analgesia. *Anesth Analg* 2011;**113**:1444−9.
11. Chen HH, Livingston SA, Holland PW. Generalized equating functions for NEAT designs. In: Davier AA, editor. *Statistical models for test equating, scaling, and linking*. New York: Springer; 2011. p. 185−200.
12. Chang KY, Tsou MY, Chan KH, Chang SH, Tai JJ, Chen HH. Item analysis for the written test of Taiwanese board certification examination in anaesthesiology using the Rasch model. *Br J Anaesth* 2010;**104**: 717−22.
13. Eignor DR. Linking scores derived under different modes of test administration. In: Dorans NJ, Pommerich M, Holland PW, editors. *Linking and aligning scores and scales*. New York: Springer; 2007. p. 135−59.
14. Hu H, Rogers WT, Vukmirovic Z. Investigation of IRT-based equating methods in the presence of outlier common items. *Appl Psychol Meas* 2008;**32**:311−33.
15. Hanson BA, Béguin AA. Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Appl Psychol Meas* 2002;**26**:3−24.
16. Arai S, Mayekawa S. A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika* 2011;**38**:1−16.