Application of Many-facet Rasch Model in In-basket Tests in China

Ruosong Yao

School of Education, Guangzhou University, Guangzhou, Guangdong 510006, P. R. China

Keywords: In-basket test,Bias analysis, Many-facet Rasch model, Personnel assessment

Abstract. The purpose of this paper is to utilize many-facet Rasch model (MFRM) to examine raters' severity/leniency, internal consistency, dimension difficulty, and examinees' ability level in order to further discuss the biases in the assessment center rating. Research questions are tested on a sample of 138 examinees who participated in the in-basket test and 6 raters who rated the test in 2010. The raters were divided into trained and untrained group to finish the work independently. An MFRM analysis was conducted to test the research questions. MFRM can estimate examinees' ability value independently of all types of variances, thus providing insights into the assessment field. MFRM also provides analysis of rater's effects, assessment dimension difficulty, and bias in assessment center test.

Introduction

The empirical study of assessment center (AC) was first conducted by British military psychologists and it has been applied in human resource management for more than 50 years[1] . In the mid 1980s, human resource assessors in China first introduced simulation exercises in order to select middle and top managers in state-owned enterprises. Despite the wide acceptance of assessment center method, questions arise concerning the integration and interpretation of the rating score. In China, the total assessment center test score is either a weighted composite of every dimension rating or by means of structural equation analyses, or find validity of AC within the exercises. These approaches have stemmed from classical test theory (CTT). They focus on both dimensions and test method, but ignore the variances brought about by the raters and ratees, thus making it difficult to identify the biases that emerge in the rating process.

There are several deficiencies and inconveniencies while just using CTT for assessment center test analysis. They include rater effects analysis (severity/leniency effect, central tendency, halo effect, and training effect), dimension difficulty analysis, and examinees' ability analysis.

The severity /leniency effect refers to the problem that occurs when a rater tends to do all the rating either harshly or leniently. Previous studies indicate that raters' severity or leniency can influence the rating process but their findings vary[2] [3] . With the advancement in personnel measurement, researchers turn to modern psychometrics to identify the root cause of errors and make adjustments to accommodate different types of errors.

Compared with classical test theory (CTT), MFRM, which incorporates more facets (e.g., examinees, raters and items) into the analysis, could provide more detailed information on the basis of the measurement of different facets [4] [5]. The monitoring result can be used to amend the test paper. Such approach can enhance rating accuracy in the assessment.

In recent years, with the evolution of modern psychometric theory, researchers are not satisfied with just developing simulated tests. Examining the essence of simulated assessment becomes their overriding concern[6] (Lance, 2008). Are there any mistakes in the raters' rating? How about the task difficulty?

Research purpose

Our study aims to investigate three research questions:

Question 1: Can we identify a "problematic" examinee who has a relatively large difference in rank ordering by comparing the decision made on the basis of FACETS estimation of examinee's ability measure with the decision made on the basis of examinee's total score?

Question 2: Does the research allow us to discern the rating quality of raters by analyzing raters' severity or leniency and internal consistency?

Question 3: Based on bias analysis result, are we able to detect biases displayed by "problematic" examinees and unqualified raters so as to analyze the underlying cause of the bias?

Method

Instrument

We developed the in-basket test and the corresponding assessment criteria. The profile of in-basket test used in the assessment is composed of four dimensions: organizing and planning, communication and coordination, problem solving, and holistic view.

Participants

There are 138 students in their senior year from Guangzhou university, who participated in the in-basket test in 2010. All participants, with an average age of 21. The proportion of male and female participants was 1:3. Six raters were divided into trained and untrained groups, with raters A, B, C in the untrained group. Prior to the rating, a 40-hour training session was run for three raters in the trained group.

Procedures

All participants must accomplish the in-basket test within the allotted time. Six raters then rated ratees' responses to the tasks with a 10-point rating scale.

Result and analysis

Result of analysis of ratee's ability

Table 1. Some ratees' total score, ability estimates by MFRM and ranking order

Examinees	Observed Average	Measure	S.E.	Infit MnSq	Outfit MnSq	Ranking order	Ranking based on total score	Ranking variance
14	6.58	0.59	0.17	0.58	0.57	10	11	-1
19	2.38	-2.46	0.21	0.71	0.72	138	138	0
20	4.71	-0.64	0.16	0.25	0.27	68	63	5
22	4.83	-0.56	0.16	1.14	1.10	59	80	-21
50	6.58	0.59	0.17	1.29	1.32	11	10	-1
59	7.46	1.28	0.19	2.31	2.21	2	1	-1
134	5.08	-0.39	0.16	0.78	0.77	42	39	3

Note：RMSE: 0.17 Adj S.D.:0.79, Separation: 4.58; Separation reliability: 0.95

In Table 1, column 1 represents examinees' number and column 2 is the mean observed value, indicating the average score of four dimensions that the raters rated the examinees. Column 3 is the examinees' ability estimates assigned by MFRM. Column 7 represents the ranking order based on the ability estimates by MFRM and column 8 is the ranking based on the total score. The last column shows the ranking difference between MFRM and the total score.

As Table 1 illustrated, the ranking of ratees would undergo changes when different statistical methods were used. Examinee 22 ranks 80 when total score is considered, whereas 59 in ranking order by MFRM. The ranking difference is 21, which exhibits the largest difference in ranking among all the examinees. 55.80% of all the examines encountered the ranking variance, indicating that different statistical methods can influence the assessment result. MFRM analysis shows that examinees' ability estimates ranged from -2.46 Logits to 1.28 Logits, with the mean ability value of -0.69 Logits. There are 41 examinees that have an *Infit* MnSq value of more than 1.2, equivalent to

29.71% of all the examinees. This shows that raters rate these 41 examinees diversely and there exists rating bias. In particular, Examinee 20 gets the lowest Infit MnSq value, indicating a lack of variation among raters' rating and a central tendency effect. In addition, from Table 1 we can see that the separation index is 4.58, with its reliability of .95, indicating that the ability level of examinees is significantly different. This was confirmed by the chi-square of 2682.2, with statistical significance at $p<.01$（$\chi^2_{(137)} = 2682.2$，$p<.01$），indicating that examinees' ability varied significantly from one another.

Result of analysis of raters' severity/ leniency and internal consistency

Table 2. Report of raters' severity/ leniency and internal consistency

Raters	Observed Average	Fair-Average	Measure	S.E.	Infit MnSq	Outfit MnSq
A	5.90	5.98	-0.84	0.04	1.17	1.22
B	3.95	3.83	0.49	0.04	1.12	1.15
C	5.54	5.58	-0.60	0.03	0.83	0.81
D	4.24	4.14	0.28	0.04	0.97	0.97
E	4.22	4.12	0.29	0.04	1.18	1.17
F	4.13	4.02	0.36	0.04	0.74	0.73

Note: RMSE: 0.04; Adj S.D.: 0 .51; Separation: 14.44; Separation reliability: 1.00

Table 2 is a report of 6 raters' scores by means of MFRM. The result provides information in regard to rater's severity/leniency measures. Column 4 in Table 2 shows rater's severity estimates. The larger the value, the stricter the rater is. On the other hand, the smaller the value, the more lenient the rater is. The severity measures fluctuate around 0 Logits. From Table 2, we can see that Rater B was the strictest, whereas Rater A was the most lenient. Their variance was 1.09 Logits. Moreover, Rater C had a negative severity estimate, indicating that Raters A and C exhibited considerable leniency in the rating. According to the fit, except for Rater F, all other raters' Infit MnSq values ranged from .84 to .99, indicating a good internal consistency and a good management of severity issues. Infit MnSq measure of Rater F is less than .8, less than the expected value assigned by the model. This illustrates that Rater F took a conservative approach to rating and the chance of his giving a high score is limited. As for the chi-square test, the concomitant probability is less than 0.01（$\chi^2_{(5)} = 1271.4$，$p<.01$), indicating that six raters were not equally severe.

Result of analysis of dimension rating

As shown in Table 3, in terms of the estimates of difficulty of the criteria measured, *holistic view* is the most difficult while *communication and coordination* is the easiest. The Infit MnSq value shows all dimension rating values approximately fit the model with the exception of problem-solving dimension's value slightly lower than 0.8, which was still within the acceptable range. The chi-square test showed that the concomitant probability is less than 0.01（$\chi^2_{(3)} = 409.2$，$p<.01$), indicating there are significant differences in the difficulty of the criteria measured.

Table 3. Analysis of dimension rating

Criteria	Measure	S.E.	Infit MnSq	Outfit MnSq
communication and coordination	-0.31	0.03	1.13	1.15
holistic view	0.49	0.03	1.12	1.11
organizing and planning	-0.14	0.03	0.98	0.97
problem solving	-0.04	0.03	0.77	0.79

Note: RMSE: 0.03; Adj S.D.:0.30; Separation: 10.21; Separation reliability: 0.99

Result of analysis of the rating scale

With an overview of the number of times that six raters used the scale, we can see that a bulk of raters centered on scale levels 3-5, with scale level 4 topping the list and followed by level 5. This demonstrates, to some extent, a central tendency. The mean ability values of examinees stands for the relationship between the scale level and examinee's ability. It is generally acknowledged that examinees' ability measure should increase with the progression in the scale level number.

Discussion

Examinees' ability estimates

Previous assessment calculates ratees' total score based on the summation of raters' dimension rating of ratees. However, MFRM can estimate examinees' ability value independently of all types of variances, thus providing new insight into the assessment field, that is, selecting examinees according to their ability level. Take the in-basket data of our study as an example. Assume that only top 10 examinees can be selected according to the total score that they got. It is obvious that Ratee 50 will be selected and Ratee 14 will be rejected. The largest variance in ranking order was found on Ratee 22. Therefore, it is suggested that the top-ranking examinees be under more thorough scrutiny before being selected and test administrators should avoid the final decision-making solely on the basis of ratee's ranking order.

Rater effects on the assessment

Raters who display considerable harshness or leniency can affect ratees to some extent. The more severe the rater, the less likely it is for the ratee to get a high score. MFRM gives specific measure logit to rater's severity or leniency, an objective index that helps differentiate raters.

Internal consistency is an index that measures whether the same rater rates the examinees in a consistently stable way. Classical test theory is concerned with the extent to which raters are consistent on the group-level and is by no means to measure the internal consistency of individual raters. On the contrary, MFRM provides analysis of intra-rater consistency. By analyzing internal consistency in relation to different aspects of the rating situations such as examinees or rating dimensions, MFRM helps test administrators detect, monitor, or train those raters who are not internally consistent.

MFRM analysis of the assessment dimension difficulty

Subjective assessment techniques demand rater's assessment on the basis of observations of examinee's behavior and that different dimension rating affects examinees' total score. The result of our study illustrated that holistic view was the most difficult dimension in rating. Holistic view refers to the extent to which an individual can objectively perceive his or her role in the job and in the organization as a whole and can effectively manage the work and time. The dimension being severely scored may be partly due to the fact that the ratees are seniors in college and they lack working experience in the real world. By assessing the dimension difficulty and doing the competency-based job analysis, test administrators can assign a dimension difficulty score to each job and then select the examinees for the job.

Conclusion

With FACETS software, the above MFRM analysis of the present in-basket test has yielded three major findings as follows:

A). MFRM can produce an estimated ability value in line with the examinee's actual ability. Thus, it provides comprehensive and accurate information for personnel selection;

B). MFRM analysis can discern the rater effect such as severity/leniency, central tendency, and training effect in rating process. The statistical analysis based on MFRM manages to decrease rater biases and improve rating accuracy by detecting raters who are significantly unstable in internal consistency or who are overly severe or lenient.

C). Variance exists between raters and examinees and significant variance is found between raters and dimensions. This demonstrates the necessity of special training for individual raters and a revision of rating criteria for some dimensions. Therefore, the analysis of the rating dimension can result in a good match between the rating dimension and the assessment purpose.

As a result, the MFRM output can assist test administrators in the rating process. Test designers can then put forward corresponding measures to improve the assessment quality. This validates the unique advantage item response theory (IRT) that has in the analysis of subjective assessment result in assessment center. Therefore, we expect that by combining classical test theory, IRT could help researchers analyze different traits of the assessment result in a more comprehensive manner, thereby

improving the rating accuracy, providing timely feedback, enhancing assessment techniques and raters' rating method, and improving the application accuracy of assessment center method in personnel assessment.

References

[1] Thornton III, G. C., & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. Human Resource Management Review, 19(3), pp. 169-187.

[2] Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), pp. 399-418.

[3] Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4), 371-389.

[4] Basturk, R. (2008). Applying the many-Facet Rasch model to evaluate powerpoint presentation performance in higher education. *Assessment & Evaluation in Higher Education*, 33(4), pp. 431-444.

[5] Wolfe, E. W., Myford, C. M., Engelhard Jr, G., & Manalo, J. R. (2007). Monitoring reader performance and drift in the AP English literature and composition examination using benchmark essays (pp. 1-39): College Board Research Report.

[6] Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology*, 1(1), pp. 84-97.