

Many-Facets Analysis of Overseas Students' Evaluation of Teaching

Gao Qinghui

Overseas Education College,

Xiamen University

Xiamen, China

E-mail: gqh@xmu.edu.cn

Abstract—The study employed Many-Facets Rasch Model (MFRM) to validate Overseas Students' Evaluation of Teaching. 5 teachers and 43 foreign university students were involved in the study and five categories of the rating scale were employed to score the teachers' teaching quality of TCSL (Teaching Chinese as a Second Language). Results show that the rating scale was able to distinguish between different levels of teaching quality, there were significant differences in the teachers' teaching quality of TCSL (Teaching Chinese as a Second Language), and that raters used the scale in a consistent and consensus way.

Keywords—Item response theory; foreign students; Evaluation of Teaching; many-facets Rasch model;

I. INTRODUCTION

As more and more overseas students study in China, many schools adopt the method of "students' evaluation on teaching" in order to ensure the quality of education to international students. Specifically, students make evaluation on teachers' teaching content, teaching method and teaching effect, and then the school administrative department makes the evaluation on teachers according to the evaluation results. As such an evaluation is closely related to the teachers' interest (such as appointment, promotion, etc.) and helps improve teaching quality, some scholars think that the evaluation on teaching is "a fundamental system of ensuring the teaching quality of the whole school". However, studies have shown that the teachers' score in the evaluation of teaching is not entirely in the positive correlation with teaching quality of teachers and that even there exists "reverse evaluation", specifically, those teachers who cater to students, treat their students in a loose manner and are not strict in their teaching score high while those who are strict in each teaching step score low. In fact, studies have pointed out that students' evaluation on teachers has something to do with many factors, for example, students' enthusiasm and interest. Typically, the teachers who teach difficult courses relatively score lower [1].

In the typical teaching evaluation, the evaluation on the teacher is based on the comparison result after all the students' scores for individual teachers are added up to. However, we cannot know from the scores of teachers whether students' evaluation are too strict or too loose and whether students can stick to their own evaluation criteria during the whole evaluation process to give a reasonable and fair score for every evaluated teacher. McNamara (1996) [2]

pointed out that raters might be very strict or very loose with some group of (or one) examinees or they might be very strict or very loose with some criterion. In addition, as foreign students from different countries have quite different cultural backgrounds and Chinese learning motives, can their evaluations on teachers be used as an indicator of judging teaching quality, or to what extent their evaluations reflect the teachers' true teaching status, or will these problems exist in the evaluation? To solve these problems, this article attempts to analyze the evaluation results with Many-Facets Rasch Model (MFRM).

II. RESEARCH METHOD

A. Data

The raters for the paper are 43 international students from XX University, who are required to make evaluations on five teachers (respectively marked as S1, S2,...,S5). Four categories of the rating scale were employed to score the compositions of these students. It is a five-point scale (4=very good; 3=good; 2=reasonable; 1=bad; 0=very bad) and five categories, Earnest and responsible teacher, patient and friendly (W1); Knowledgeable teacher, clear and easily understandable explanation (W2); Sufficient and effective class activities (W3); Interesting and enlightening teaching (W4); Teaching is in combination with Chinese culture (W5). 43 foreign students are in their third years and from the same class of the university. The experiment was made in the way of the anonymous questionnaire survey. As the questionnaires were handed out and collected in the classroom, the collection rate reaches up to 100%.

B. Many-Facets Rasch Model (MFRM)

Many-facets Rasch model is an extension of the basic Rasch one-parameter item response theory (IRT) model. IRT is a class of psychometric models used to estimate examinees' ability and the difficulty of test items on the same scale [3].

The basic Rasch model includes two facets of examinees and items. It completely depends on the ability of examinees and difficulty degree of items to deduce the probability formula of right answers. MFRM extends the basic Rasch model by adding parameters describing facets of measurement interest other than item difficulty (such as rater severity or task difficulty) to the model [4]. MFRM attempts to free each examinee's measure from the effects of

differences in rater severity or task difficulty [5], and the estimated value for each parameter is logits so as to remove influence of various factors over the ability of examinees in the subjective evaluation and increase the reliability of the result. The paper applies the model to the evaluation of foreign students' writing. The following form is employed [6]

$$L_n \left[\frac{P_{ijk_x}}{P_{ijk_x-1}} \right] = B_i - D_k - C_j - F_x$$

P_{ijk_x} = the probability of examinee i being awarded a rating of x when rated by rater j on task k ;

P_{ijk_x-1} = the probability of examinee i being awarded a rating of $x-1$ when rated by rater j on task k ;

B_i = the teaching quality of examinee i ;

D_k = the difficulty of task k ;

C_j = the severity of rater j ;

F_x = the difficulty of rating threshold x , relative to rating threshold $x-1$

Model construction and data analysis was carried out in FACETS version 3.68.1. There are 3 facets: evaluated teachers (marked as examinees, 40; s1-s5); raters (marked as rater, 43; R1-R43); evaluation criteria scale (criteria, 5 items, w1-w5).

III. RESULT EXPLANATIONS AND DISCUSSION

A. General Reports

Figure 1 was an overview of the 4 facets. Column "Measr" showed the logit measures for examinees. Column "examinees" showed the examinees' teaching quality estimates. The top of the scale indicated highest teaching quality; and the bottom of the scale indicated lowest teaching quality. It decreases progressively from top to bottom; it can be seen from Figure 1 that the student s1 is the best in teaching quality (above 4logits) while the s4 is the worst (below 1logits). Column "raters" was the rater severity estimates. The higher the rater was located on the scale, the more severe he tended to be in the rating. The measure above the 0.00 logit indicated a more severe rater, and the measure below the 0.00 logit indicated a lenient rater. It can be seen in the illustration that R16, R19, R23, R9 and R43 are stricter, whose logit scales are above 2logits. Among of them, R43 is the most vigorous. In addition, R11, R15, R2, R32, R7 and R6 are looser, whose logit scale are below -2logits. Column "criteria" shows the difficulty degree of each item, that is, the difficulty to get high score in the item; the one which is marked high is more difficult. The most difficult one is W4 and the easiest one is W1. Maybe the overseas students have higher requirements for vividness in the teaching, so Chinese teachers are hard to get high scores. w1 (whether the teacher

is responsible or patient or friendly) has the lowest difficulty, which shows that most teachers can meet this requirement and get a high score. Finally, column "scale" shows the 5-point rating scale and the distance between each step on the scale possibility of scores teachers can get. It indicated the scores that examinees at a certain ability level on the scale were likely to receive. For example, teachers whose ability level is above 0 logits but below 1logits had a probability of receiving a score 3, etc. From figure 1, enables us to intuitively learn about the basic information of overseas students' evaluation on teaching, each facet is analyzed next in order to get more explicit information.

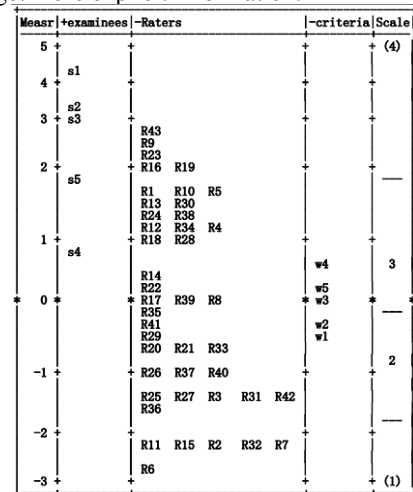


Figure 1 All facets Vertical Rulers

B. Examinee Reports

See Table I for each teacher's teaching quality. Column "Measure" shows corresponding teaching quality (logit is the unit). Infit MnSg (information-weighted mean-square fit) is a statistical index describing the extent to which an examinee's ratings are in agreement with what the model predicts. Infit mean squares showed the size of the randomness, i.e. the amount of distortion of the measurement system. 1.0 is their expected values; it means the actual data completely fits the model. Values less than 1.0 indicate observations are too predictable (redundancy, model overfit). Values greater than 1.0 indicate unpredictability (unmolded noise, model underfit), that is, less than 1.0 indicates too little variation and lack of independence (raters may take it for granted and give the similar score to examinee), and more than 1.0 indicates too much variation. Outfit mean square has the same form as infit, but is the conventional mean-square which is more sensitive to outliers.

The case Infit=1 seldom happens. In actual application, usually a numerical interval $[\alpha, \beta]$ is adopted and the case $\text{Infit} \in [\alpha, \beta]$ is regarded as complete fit while $\text{Infit} > \beta$ means underfit and $\text{Infit} < \alpha$ means overfit. In the paper, according to Wright & Linacre [7], $\alpha=0.7$ and $\beta=1.3$. It can be seen from Table I that there is a significant difference among the teachers' teaching quality. The separation index is 8.13, its reliability is up to 0.99, indicates that the assessment can

effectively separate teachers according to their level of teaching quality with a high degree of confidence.

According to the fit, only s2 (fit Mnsq=1.5>1.3) is underfit, which shows that the students' evaluations on s2 are very inconsistent. The further interviews with students tell that s2 has rich knowledge, is serious in teaching and is very strict with students. As most students hope to gain more knowledge, they give high scores for this teacher. However, for those students who take the obtainment of the diploma as their only goal, they are dissatisfied with his/her strictness in the study, so they give low assessment. It can be seen that there exists "reverse evaluation" in the teaching evaluation of overseas students. However, such students are in the minority, so s2 still gains good evaluation of 3.21logits, who ranks second.

TABLE I. EXAMINEE FACET REPORT

Exam	Measure	Infit MnSq	Outfit MnSq
s1	4.42	0.21	1.02
s2	3.21	0.15	1.55
s3	3.00	0.15	0.86
s5	1.80	0.11	0.80
s4	0.81	0.10	0.99
Mean	2.65	0.15	1.04
S.D.	1.24	0.04	0.27

Note : Separation 8.13 Reliability .99
chi-square: 375.8 d.f.: 4 significance (probability): .00

C. Rater Reports

TABLE II. RATER MEASUREMENT REPORT

Rate	Measure	Infit MnSq	Outfit MnSq
R43	2.53	.28	1.25
R9	2.29	.28	.68
R16	1.97	.29	.96
R13	1.44	.31	.34
R24	1.35	.31	2.07
R38	1.35	.31	.34
R18	.94	.33	.63
R39	.03	.40	.81
...
R41	-.31	.43	3.39
R42	-1.28	.57	.98
R36	-1.64	.64	.85
R6	-2.90	1.04	.89
Mean	.00	.46	1.00
S.D.	1.47	.18	.57

Note: Separation 2.83 Reliability .89
chi-square: 435.7 d.f.: 4 significance: .00
Inter-Rater agreement opportunities:
Exact agreements: 53.0% Expected: 52.6%

When the overseas students are taken as the rater, their personal characteristics have a great influence on the results of teaching evaluation. It can be seen from Illustration 1 that there exist great differences in their strictness of evaluation.

From Table II, we can get more information after further analysis. (Partial data is omitted to save space). It can be seen that the overseas students vary greatly in the strictness when they are required to evaluate the teaching of the teachers: R43 is the strictest, whose logit scale reaches up to 2.53logits; R6 is the least strict, whose logit scale is -2.90logits; The difference between the strictest and the loosest one is 5.43. The separation index and reliability are 2.83 and 0.89 respectively, which means there is statistically significant difference among raters. In terms of the chi-square test, the concomitant probability is 0.00. As for the consistency between raters, the exact agreement is 53.0%, greater than the expected value of the model, 52.6%, so it is acceptable.

From the "Infit MnSq" column of Table II, we can see the consistency of overseas students in evaluation on teaching. For example, the infit of R24 is 2.07 and the infit of R41 is 3.39, which are beyond the scope of $\text{infit} \leq 1.3$. This is the case of "underfit", which shows that these overseas students have subjective instability in evaluation on teaching. For R13 and R38, the infit is 0.34, which is less than 0.7. This is the case of "overfit", which shows that they give similar or the same scores and evaluations for different levels of teaching as they don't distinguish the teaching level when they make the evaluation on teachers. From the further analysis of questionnaires, it can be found that this phenomenon has much to do with overseas' students' cultural background, Chinese language proficiency and other factors. For example, the students from Japan and South Korea have the tradition of respecting teachers, so they prefer to give high evaluation to each teacher; The students from Indonesia often have higher Chinese level and they cherish higher expectation for their own learning results and teachers' teaching level, so they tend to give lower scores to the teachers. Although we emphasized objectivity of evaluation through mobilization before arranging the task of evaluation on teaching, the analysis shows that there still exists the case of extreme strictness or extreme looseness. Therefore, in the future, before the evaluation on teaching is made, the interview should be made and some proposals for evaluation on teaching should be put forward purposefully to minimize the bias in the evaluation on teaching. Of course, generally speaking, in this evaluation, the overall consistency of the raters and their own consistency meet the requirements, which can be used as a basis for evaluation.

D. Bias Analysis Results

In order to further understand whether some overseas students deviate from the principle of impartiality because of their special liking or hatred for some teacher or specific evaluation item, we have to make the deviation analysis.

First, the interaction between overseas students and the evaluated teachers must be considered. There are 5 teachers who are evaluated and 43 overseas students participating in the teaching evaluation, so there is a total of 215 (5×43) items. In Table III, 10 items with obvious deviation are listed, which have the absolute value of t more than 2 and concomitant probability of $p < 0.05$. the "Obs. Score"

columns show the total points the corresponding overseas students score for this course. The “Exp. Score” column shows the expected total points of the model. The “bias size” column shows the deviation size. When the observed value is less than the expected value, the deviation is negative, which shows the rater is so strict that he/she scores extremely low; but if the deviation is positive, it shows that the rater is too loose. It can be seen from Table 4 that the rest are too harsh except that R16 is too loose in the evaluation on s4. We said earlier that s2 is a teacher who treats teaching seriously and carefully and is strict with students. It can be considered that r10, r24 and r41 don’t expect to be under the tutelage of strict teachers. Especially for R41, the deviation reaches -3.83 logits (It is the deviation with the largest absolute value in the total item; the p-value is as small as two-ten thousandths), which is really artificial deviation. As the evaluation on teaching is made in an anonymous way, we cannot learn more about the reason for deviation (for example, whether this student gets quite low score in this teacher’s course), but it can be considered that R10, R24, R41 are the students giving “reverse evaluation”. Although S1 is given low assessment for three times, he/she still maintains the highest assessment. Although s4 is given higher assessment once by R16 and lower assessment once by R35, the results aren’t affected. Although s3 is given low assessment twice, the deviation is not too large enough to have a big impact. In general, there are 10 items with obvious deviation in the total of 215 interacted items, accounting for 4.7% (less than 5%), which is in the acceptable range.

TABLE III. BIAS/INTERACTION ANALYSIS SPECIFIED BY: EXAMINEES AND RATER

Rate	Exa.	Obsvd Score	Exp. Score	Bias Size	S.E.	t	Prob.
R14	s1	17	19.5	-2.32	0.74	-3.14	0.035
R22	s1	17	19.6	-2.6	0.74	-3.52	0.024
R21	s1	18	19.8	-2.9	0.84	-3.45	0.026
R10	s2	13	16.9	-1.58	0.57	-2.78	0.05
R24	s2	9	17.4	-3.09	0.58	-5.33	0.006
R41	s2	12	19.3	-3.83	0.55	-6.93	0.002
R43	s3	9	14.4	-1.7	0.58	-2.94	0.043
R2	s3	18	19.8	-2.88	0.84	-3.43	0.027
R16	s4	15	9.2	1.87	0.63	2.97	0.041
R35	s4	11	15.6	-1.55	0.55	-2.84	0.047

Likewise, the interaction between raters and test items can be studied. Among all 215 items, there is only one item, that is, the concomitant probability of R11 for W1 is less than 0.05, It can be considered that there is no obvious deviation when overseas students use the evaluation criteria for evaluation.

IV. CONCLUSION

Based on above statistical output from Facets, conclusions can be drawn as follows:

Through the overseas students’ evaluation on teaching, the teaching quality of teachers can be distinguished, that is, the result of evaluation on teaching is effective. The separation index of the teaching quality of the teachers is 8.13, and its reliability reaches up to 99%.

The overseas students’ strictness in the evaluation on teaching varies, with the difference up to over 5 logits. There exists “reverse evaluation” in this activity of evaluation on teaching, that is, some students give very low scores intentionally, but they are quite few. This shows that it is necessary to illustrate the significance of evaluation and importance of fairness in evaluation for overseas students before making evaluation on teaching, just like for Chinese students. In addition, it also shows that the application of the Rasch model enables the teaching evaluation to be more accurate. If we simply add up to all the scores of the students in the teaching evaluation, the result will be unfair. For most of overseas students, their self-consistency in the evaluation meets the requirements. In addition, the consistency of evaluation between the overseas students is also acceptable.

The adopted four-level scale can truly reflect the teaching level of teachers. It is quite few that the evaluation result is influenced by some overseas student’s special view of some item. The designed criteria and the adopted scale can meet the requirements of evaluation on teaching.

ACKNOWLEDGMENT

The study is subsidized by the key project (item number DDA110200) of national scientific education plan of Ministry of Education of China.

REFERENCES

- [1] Marsh,H.W. Roch E.L.(2000). Effects of grading leniency and low workload on students’ evaluations of teaching:Popular myth, bias,Validity, or innocent by standers?.*Journal of Educational Psychology*,92(1),202–228.
- [2] McNamara , Measuring second language performance. London: Longman. 124-125.
- [3] S.M. Downing and T M.Haladyna, Test Item Development: Validity Evidence from QualityAssurance Procedures. *Applied Measurement in Education*, 2003, 10(3): 61-82 .
- [4] J M Linacre and B D Wright, Construction of Measures from Many-facet Data. In Smith, EV, Smith R M. *Introduction to Rasch Measurement:Theory, Models, and Applications*. Maple Grove, MN: JAM Press, 2004, 66-167.
- [5] J.M. Lionacre, Predicting responses from Rasch Measures. *Journal of Applied Measurement*, 2010, 11(1).
- [6] Lunz,Wright and Linacre. Measuring the impact of judge severity on examination scores. *Applied Measurement in Education* , 3, 331–345, 1990.
- [7] B D Wright and J M.Linacre, Reasonable Mean-square Fit Values. *Rasch Measurement Transactions*, 1994, 8(3): 370