Research on TEM-4/8 and IELTS Computer Rating Model Based on Rasch Model

Hu Jianhong

Wuwei Occupational College, Wuwei, Gansu 733000, China

Keywords: Rasch Model, TEM-4/8, IELTS, Computer Rating Model

Abstract: Based on the multi-level Rasch model, this paper compares and the scores of college teachers' scores in English composition and the quality of peer assessment, and the feasibility and necessity of introducing peer evaluation in writing teaching. But overall, the scoring results of the two types of scorers have a high degree of consistency, thereby effectively improving the scoring effect, enriching teaching methods, and enhancing the effectiveness of English writing teaching. This paper evaluates the construct validity of the project by investigating the underlying structure of the TEM-4/8 and IELTS grammar vocabulary project and the factors that influence the construct. The expert judgement method was used to collect the evidence of content validity. The Rasch model the effect of the response of the title question on the validity of construct validity, and the confirmatory factor analysis was used to understand the construct of the project. The results of the study show that the contents of the test sites are very consistent; most of the topics are fitted to the model, and the project has good one-dimensional characteristics. The answers to the questions are basically independent of each other. However, the project is more difficult than the candidate's ability. Lead to a few difficult topics that do not fit the model and contribute insufficiently to the construction of the project; Grammar and vocabulary belong to two very relevant but independent constructs

1. Introduction

Writing is an important part of the English teaching system, but it is characterized by high input and low output. The reason is that in the traditional teaching and learning model, students passively accept knowledge as objects, have a low degree of participation, and are not enthusiastic. In terms of assessment methods of English writing alone, the traditional English writing teaching evaluation system is too single. In the classroom or test environment, teacher scoring is generally used as the main or only mode of writing evaluation. Teachers often have to spend a lot of time and energy to evaluate student writing. Many students have failed to read and digest the teacher's review carefully, which ultimately results in little teaching effect. It is difficult for students to effectively improve their English writing ability. In view of this, for the English writing scores, it is necessary to further improve the status and participation of the students or the participants in the writing assessment. Evaluation is an important part of the learning process (Johansson S, 2014; Mahoney E R, 2015). A reliable score can provide diagnostic feedback for teaching and provide a basis for daily teaching decisions. Allowing students to participate in the evaluation process and using peer review as an important supplement to the English writing teacher's score will help improve students' autonomy. Not only can they change their passive role in evaluation, but also allow them to evaluate other people's compositions. The purpose of this study is to compare and analyze the scores of teachers and student peer assessments by quantitative methods, and to explore the feasibility of introducing peer review in college English writing courses. Foreign Research Foreign researchers have conducted extensive research on assessment of foreign language writing.

Teacher evaluation is considered to be the most basic component of the writing process. It has always occupied a central position in the field of evaluation of second language writing. However, with the development of higher education, the research focus of foreign researchers on the assessment of foreign language writing has shifted to alternative ways of composition assessment (Ghazzal, 2014). Different forms of scoring, especially peer review, have been introduced into

  DOI: 10.25236/iwedss.2019.310

college English writing classes. The study found that: There was no significant difference between teacher's score and student's peer assessment; peer review results often have a high correlation with teacher's assessment results. In addition, peer evaluation can not only allow students to score each other, improve the efficiency of writing scores in the classroom environment, but also help students gain valuable experience in assessing composition. The purpose of this study is to compare and analyze the scores of teachers and student peer assessments by quantitative methods, and to explore the feasibility of introducing peer review in college English writing courses. This study aims to further investigate the validity of the grammar vocabulary project(ElifKantarcıoglu, 2010; Aryadoust, 2011). The research questions mainly include two aspects: (1) The contribution and influence of the survey topic on constructs; (2) The establishment and evaluation of the construct structure of the project, using Rasch separately. This study first briefly reviews the study of grammar testing and vocabulary testing, and then uses Rasch to analyze the attributes of topic and grammar vocabulary items on the basis of content analysis, and finally establishes the initial TEM-4/8 and IELTS grammar vocabulary projects based on content analysis. The model was revised based on the results of content analysis and Rasch analysis to understand the contribution of the topic to the factor structure.

2. Research Backgroud and Rasch Thery

2.1 Computer scoring system

Computer scoring is a subjective scoring. The credibility of the score is influenced by various factors. The scorer is an important factor influencing the score result. Whether the scorer can maintain consistency in the scoring process affects the fairness of the scoring. Due to the nature of the writing assessment, the subjective judgments in the scoring are unavoidable. Therefore, many factors may cause different scoring results, such as the background and experience of the scorer's work, the individual's understanding and mastery of the scoring standards, and the expectation of the subject. And the views, attitudes, etc. on the exam. The may unconsciously bring a number of personal factors into the established grading standards (Baghaei, 2015). Research indicates that the grading staff may have the following differences: 1) The overall width and severity of the scale are different; 2) A certain group of candidates are more or less stringent; 3) In some aspects, such as writing and grammar; 4) understanding and applying scoring standards, different ranges of points; 5) consistency of individual points. At the same time, the scorers have a big difference in the severity of the score, and it is difficult to maintain consistency in the composition score.

The Rasch model was created by the Danish mathematician Georg Rasch and is the application of project response theory in practice. It is a single-parameter project response theory model, which is often used to the difficulty of multiple-choice test questions and the ability of candidates. The multi-layer Rasch model is an extension of the Rasch model and is suiTable for non-machine operation. It is also often used in the analysis of subjective test questions. It is possible to the ability of candidates on subjective test questions on the same Logit scale, the difficulty of test papers, the strictness of the test teacher, and the accuracy of the rating scale. The performance of other aspects and the interactions between them also help to determine whether there are significant differences between the various aspects of the composition, such as the difference in the ability of the candidates, the severity of the assessment, and so on. In addition, it can also perform Rasch model fit analysis for each aspect (Varley , 2011). Many studies use this model to application tests such as writing and speaking. Computer software "FACETS" can explain the rating of the scorer writing test. This model is suiTable for verifying the consistency among scorers, the differences in the severity of the scores, the difference between the scorers' understanding and application of the scoring standards, and the effect of the scorer's errors on the scores of the individual candidates; it is particularly suiTable for testing the scoring and can be tested The results are fed back to the scorer for improvement (Daud, 2016). The model also has a distinct advantage, that is, it can evaluate the real ability of the subject by the difficulty of the test. Using this model to the scoring can obtain information about the interaction between the subject, the subject, the question, and the scoring

criteria, respectively, and how to understand the test result. , how to improve the organization of the score, etc. are also helpful. The English test score is a subjective scoring, as shown in Figure 1.

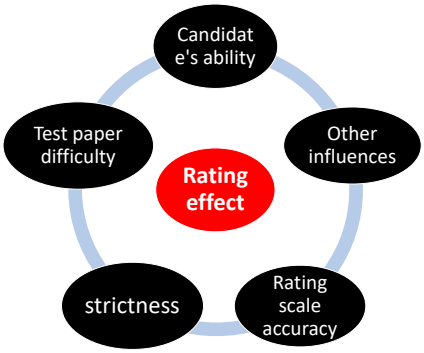

Figure 1. Subjective scoring

2.2 Research methods

One hundred and thirty-four subjects in the second grade of an ordinary middle school in the same city were selected as subjects for this study, including 150 girls, 154 boys, 143 science students, and 161 liberal arts students. The school's English scores are moderate in the city. Students come from a variety of sources and have a certain degree of representation.

Test Tools The materials used in this study were the quiz fill-in questions for the 2016 national college entrance examination TEM-4/8 and IELTS. Therefore, their attention to the topic of Volume II is not high and the possibility of prior exposure to these questions is low. 2.3 Procedures The study consists of three main steps: 1) Collect personal information of participants, such as name, class, gender, and major 2) Participants complete the 2016 national college entrance examination within the required 20 minutes. The TEM-4/8 and IELTS quiz questions are completed by the cooperating teachers. 3) Collect the test answer and count the data. You get 1 point for the correct answer and 0 points for the error. In order to better understand the characteristics of the problem, the author first the questions and test points, and puts the collected data into the Rasch model for analysis. 2.4 Data statistics model, study uses the software to collect the collected test data, and uses the commonly used indicators of the Rasch model to examine the 2016 national college entrance examination TEM-4/8. The Rasch model indicators used are: model fit values, error statistics, bubble charts, reliability and separation factors, variable plots, and DIF values. When the fit value is 1, it means that the data fits the model completely. Since the subject of this study is a large-scale examination, the mean-square fitting range of all questions in this question type should be within the range of 0.8-1.2. The quality of the TEM-4 8 and IELTS cloze questions of the RASCH model examined by the common indicators of the Rasch model is shown in Figure 2.

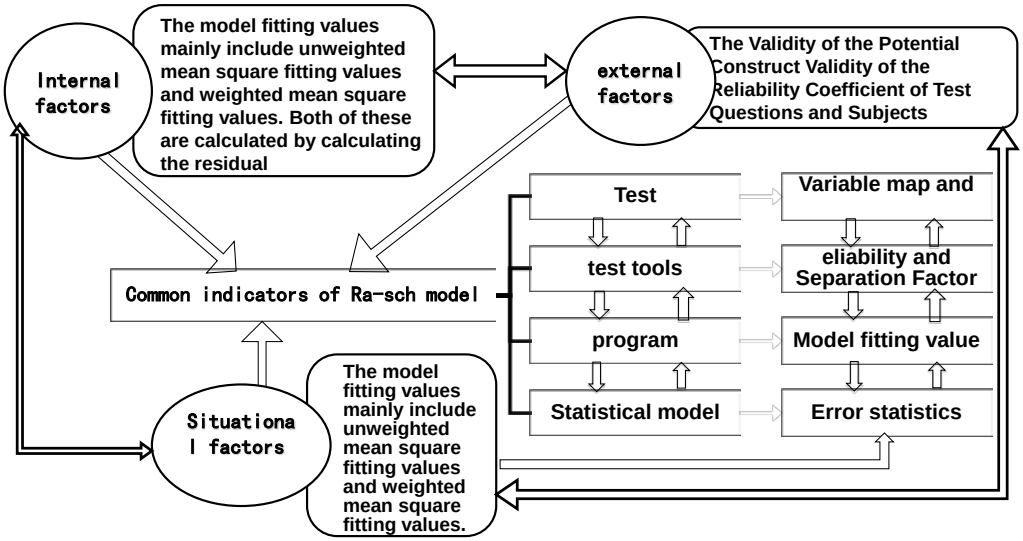

Figure 2. Commonly used metrics of Ra-sch model

3. Research Design

3.1 Research steps

The researchers randomly selected 10 essays from TEM-4/8 and IELTS 50 papers from a non-English major freshman class in a key university in Beijing. These essays were first graded by the English teacher of the class, and then the other three scorers scored the 10 essays again according to the same grading standard. In this way, each essay received 4 scores given by four scorers. The four raters are two men and two women. The No. 1 scorer is the English teacher and female of the class; the No. 2 scorer is the graduate student and the female who has no teaching experience; the No. 3 scorer is the graduate student who has three years of teaching experience in college English; the No. 4 scorer is High school teaching experience in postgraduate, male. Each rater has a scoring standard that stipulates the content and requirements of the rating. Out of 25 points, divided into five levels: 5) 20-25; 4) 15-19; 3) 10-14; 2) 5-9; 1) 0-4. Each level has a brief textual description, requesting content, language, grammar, appropriateness, and length. The scorers all scored themselves according to the standards, and did not make any marks on the papers, and then exchanged the papers for re-evaluation.

1) Textual genre and theme analysis: The quiz fill-in question for the national college entrance examination TEM-4/8 consists of 20 short-term questions, accounting for 20% of the total score of 150. The question is composed of a 250-character short passage and 20 blanks. It requires the candidate to select an appropriate answer from the four options given in each question and fill it in the blank to make the chapter smooth and complete. This test mainly tests candidates' comprehensive linguistic abilities such as vocabulary, grammar, pragmatics, discourse, and reading comprehension. Textual genre is narrative. The theme of a chapter is a story and personal experience. This type of topic is close to the examinee's life and will not cause irrelevant factor pollution scores.

2)Topic analysis: t Candidates can be divided into four levels: words, phrases, sentences, and texts. The word level refers to the candidate's answer by word cues alone; the phrase level refers to the candidate's familiarity with the test's fixed collocations to correctly answer; sentence level refers to the candidate's need to refer to the information provided in the sentence's sentence to select the answer; Levels refer to candidates who need to refer to contextual clues for correct answers. It can be seen that the four test sites have their own characteristics and are related to each other to a certain extent. Discourse-level questions are more able to reflect candidates' language abilities than those at other levels (Chen Xiaojian 2001). Based on this, the level of the TEM-4/8 cloze test subject is shown in Table 1. There is no test point involving the word level, the discourse level accounts for 65%, followed by the sentence level 30%, and the phrase level only 5%. It can be seen that the TEM-4/8 cloze test mainly tests textual language competence and is in line with the current mainstream of language testing.

3.2 Research tools and methods

The evaluation standards and contents to the students, but also design and make the corresponding evaluation tools. In most cases, these evaluation tools usually appear in the form of evaluation Table. Not only to reflect the students' sports performance evaluation Table, and evaluation should also reflect the students learning process and learning attitude scale. The student sports basic theory knowledge evaluation can take oral, written examination, knowledge contests and other forms, according to the assessment results are converted into the score of the student’s basic theoretical knowledge of sports learning. In order to simplify the examination procedures and workload, the students physical fitness test and score can be combined with school each year to implement the "Student Physique Healthy Standard" to, the student sports study physical scores and with reference to the health standard of students' physique three physical quality score conversion into a. The choice of assessment content according to the nature and characteristics of sports skills assessment project, adopt the method of quantitative or qualitative results make the corresponding evaluation Table. Physical and motor skills progress evaluation method, first of all, according to the student physical fitness and sports skills raw scores and test scores, to measurement result is

subtracted from the original scores for the progress and to progress points control Table compiled by physical and motor skills improved evaluation physical and motor skills progress performance assessment.

By using different evaluation and measurement methods, can collect different types of data and evidence. Commonly used methods are: standardized test, standard reference test, based on the results of the evaluation, student work sample, the performance of the students observation, survey and interview. Teachers need to collect the data and evidence for analysis, the formation of a student's physical learning situation analysis results, and the objective description of the current learning situation of students. To pay attention to the following questions in the analysis: coping in the appropriate team to collect data for analysis; response from the data of various assessment methods for comprehensive analysis, in order to fully describe the state of development of the students. If there is a longitudinal data, including trend analysis and if I can get the comparative data of other groups, through comparative analysis of the development of students.

Table 1. Cloze fills the test level

Question number	40	41	42	43	45
test level	D	D	S	D	21
Question number	46	47	48	49	50
test level	S	D	D	D	P

Where D is discourse, S is sentence, and P is phrase. If there is a significant gap between the difficulty distributions between the questions, it means that the test questions do not involve the individual dimensions for which the constructs are to be measured.

3.3 Data analysis

Computer programs are used to analyze the data. For the data obtained, three-sided (writing ability, stringer severity, and scorer gender) measurement models were used.

It can be seen that the RASCH model is different from the traditional scoring criteria in describing the candidate's oral language ability linearly, but objectively gives the candidate different levels of specific language to use what kind of judgment criteria. Its main advantages are: (1) the design principle is simple and can be used to formulate scoring standards for specific spoken language or writing test tasks; (2) the developed scoring standards are easy to use, especially when scored by the personnel who participate in standard setting It will achieve a better rating reliability; (3) If the scoring criteria generated for each specific task is used in the teaching situation, it can accurately reflect the students' performance and provide feedback for the students. In short, the scoring criteria should be based on the specific test purpose and the test subject, and should be based on the decision-making based on scores. There is almost no relevant research based on candidates' actual test performance data to formulate scoring standards. This study is to make up for this deficiency, and it is also Attempts and innovations in scoring criteria development methods.

The RASCH model is one of the theoretical models of the project response. The model can be used for the quality assessment of subjective test questions. It is based on a stochastic probability model and measures each individual (candidate, grader, task, etc.) on a common logit scale at each level, and calculates the estimated error for each metric, the degree of fit to the model, and each between layers. The RASCH model developed by Linacre and Wright is computer software that can analyze and interpret scorer ratings. In view of the powerful features of MFRM, more and more scholars use the RASCH model to study L2 writing scoring problems. Using RASCH model, we can get the following main analysis data. 1. Measure: The individual's scale value on the common scale. Measure the fit between each individual's actual observations and model predictions. Including Mean Square and Outfit Mean Square. The latter is more likely to be affected by data with greater differences, so the former is generally used as the basis for determining whether an individual, it means that the model prediction is in full compliance. A fitted value greater than 1 indicates that there is a random deviation between the data and the model, while a value less than 1 indicates that the difference between the data is less than the difference predicted by the model. The

scorer’s scoring behavior analysis is shown in Figure 3.

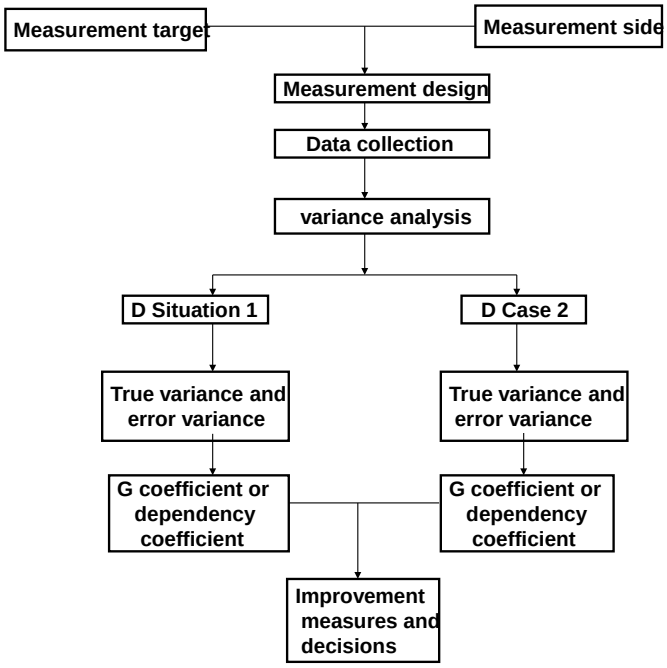

Figure 3. Grader's scoring behavior analysis

4. Result and Discussion

4.1 Overall analysis

The RASCH analysis uses a logit as the basic test unit. In the scale, 0 indicates an intermediate level. Above zero is represented by a positive number, and below zero is represented by a complex number. Matching the candidate, the greater the number, the stronger the candidate's ability; and the greater the number, the more stringent the scorer's rating.

Table 2 shows the candidates. For candidates, the higher the Rocky value, the worse the ability. From the figure, the highest Measure is 3.60, and the lowest is -2.60. It can be seen that of the 10 candidates, No. 7 candidate has the worst writing ability, while No. 1 candidate has the strongest ability. In addition, from the overall perspective of these ten candidates, the writing ability of candidates can be clearly distinguished in this sample (standard deviation 1.64, separation value 4.62, segregated index reliability 0.96, chi-square value 8.5, The level of significance is 0.39).

Table 2. Candidate Ability Table

Score	Count	Average	Fair-M	Measure	S.E.MnSq	ZStd	MnSq
4.00	23.00	23.24	3.60	0.49	0.56	3.30	2.20
78.00	4.00	19.50	19.61	1.18	0.37	54.00	0.00
10.00	10.00	67.00	5.00	7.00	34.00	4.00	12.00
76.00	4.00	19.00	19.22	0.93	0.35	7.00	1.39
9.00	9.00	7.00	5.00	6.00	15.00	9.00	1.50
75.00	4.00	18.80	19.02	0.80	0.35	13.00	0.43
6.00	6.00	17.00	26.00	56.00	24.00	6.00	56.00
68.00	4.00	17.00	16.98	-0.02	0.35	11.00	0.60
2.00	2.00	16.00	18.00	-6.00	13.00	21.00	45.00
63.00	4.00	15.80	15.54	-0.60	0.33	15.00	0.41
3.00	3.00	41.00	34.00	21.00	43.00	17.00	11.00

At the time of scoring, we hope that the scorers themselves are highly consistent and the severity of the scores is not significantly different (Linacre 1994). The 5th grader is slightly lower than 0.6, indicating that the score difference is small, but its score Its consistency is good. Overall, two of the

seven raters had poor self-stability when scored, but more than half of the raters had better self-consistent ratings. The score can be considered as more credible. Below the Table, the separation ratio is 3.38 and the separation index is 4.84, which indicates that the severity of the scorer can be roughly divided into 5 different levels; the separation reliability is as high as 0.92, and the chi-square analysis result is 76.5, P. = 0.00<0.05. It shows that the seven scorers show different severity, which is similar to the results of Ho's satisfaction, indicating that although the scorers have a good internal consistency, they generally show different severity differences. The individual severity of the rater can view the measure value and measure indicates the true severity of the rater. It is generally believed that the score measurement is greater than 0 and the score is strict; less than 0, the score is loose. From the Table, we can see that the No. 6 scorer has the best degree of severity, the No. 2 scorer is the most severe, and the No. 1 scorer is the most relaxed.

The INFIT value shows the internal consistency of the scorer. When this value is between 0.5 and 1.5, it indicates that the internal consistency of the scorer is better; when this value is greater than 1.5, it indicates that the scorer score is inconsistent; when this value is less than 0.5, it indicates that the scorer's The score is too concentrated. The OUTFIT value can indicate whether the scorer's scoring criteria are consistent. When this value is between 0.5 and 1.5, it indicates that the scorer has a good grasp of the scoring standard and has internal consistency. This also coincides with the fact that the score is too concentrated. The scorers and score criteria for the No. 2 and No. 3 scorers are relatively modest. For the No. 1 scorer, her scoring standards were relatively messy and the OUTFIT value was as high as 3.54. This may have something to do with her subjective feelings for her students.

Table 3. Grader Behavior Table

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	S.E.	Infit	Outfit MnSq	Estim. ZStd	MnSq	ZStdDiscrm
121.00	149.00	10.00	14.90	15.14	0.89	0.19	0.50	0.53	-1.20	0.61
213.00	161.00	10.00	16.10	16.16	0.38	0.23	0.61	0.29	-1.60	0.32
172.00	10.00	17.20	17.89	-0.23	0.23	0.30	0.76	-0.30	0.82	-0.10
190.00	10.00	19.00	19.49	-1.04	0.22	1.10	0.74	-0.50	3.54	3.10
168.00	10.00	16.80	17.17	0.00	0.22	0.90	0.58	-0.90	1.32	0.20
15.10	0.00	1.50	1.66	0.72	0.02	0.70	0.19	0.50	1.29	1.80
17.40	0.00	1.70	1.92	0.83	0.02	1.12	0.22	0.60	1.49	2.00

From the above results, it can be seen that the four have different levels of mastery of the grading standards and the severity of the grading. And, the model analysis results also provide gender differences in scoring. In terms of severity, male scorers are more severe than female scorers. In addition, the scorer grasps the inconsistency in the severity of the score. In addition, due to differences in the experience, experience, etc. of the four, they showed some inconsistencies in the scoring process. The specific reasons for the inconsistency of the scorer's internal scoring and the inconsistency of the scoring standards remain to be further studied.

4.2 Computer-level scoring analysis

The scores of computer scorers are different, but they are all within the accepTable range (± 1 logits) and the average width is. 00logit. Three raters were slightly more relaxed (logit value < 0), the 21 (.68logit) and 25 (.68 logit) scorers were the most severe, and the 24th raters were the most relaxed (-. 56 logit). However, the logit value difference of the strictest and loosest computer scorer is only 1. 24 logit (.68 logit ~- . 56 logit) is much smaller than the student's scorer's strictest (2.9 log log), which indicates that the computer scorer is better than the student in overall scores. Although the separation reliability (. 66) and chi-square analysis results ($\chi 2 = 14.4$, $p = .01$) indicate that there are still significant differences in the severity of computer scorer scores, the logistic values of the strictness of all computer scorers are In the accepTable range (± 2 logits), the average width is. 00 logit, indicating that the computer scorer's rating results are reasonable.

Participants with different combinations of capabilities may have different accuracy estimates. In

order to further inspect the accuracy of the selected strategy on each capability point, a grid map of the conditional estimation accuracy (AED) of each strategy in each two-dimensional capability point and a contour map are drawn to explore the number of test dimensions. The influence of the two dimensions of correlation between dimensions on RASCH, Monte Carlo simulation experiments. The two influencing factors are set as follows: (1) Number of test dimensions: There are 2 levels, 2 and 5, respectively. (2) Correlation coefficient between multidimensional capabilities: There are 4 levels in total, which are 0, 0.2, 0.5, and 0.8, which represent no correlation, low correlation, moderate correlation, and high correlation, respectively. The feasibility of the two-dimension RASCH selection strategy and the comparison among the strategies of each topic are shown in Fig.4.

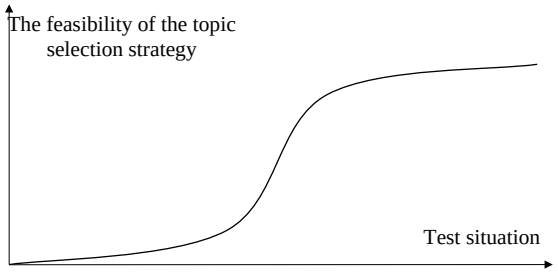

Figure 4. The feasibility of the topic selection strategy

4.3. Random effects

The random effect refers to the abnormality of the computer score in the use of some or some fractional segments, and there are obvious inconsistencies, thus showing a greater randomness. If there is a random effect, the computer score cannot distinguish the actual level of each player. From the data in Table 3, it can be shown that the player level is differentiated, and there is no random effect on the computer score in general. The random effect can also be reflected to some extent by the two-column correlation coefficient between a single computer score and other computer scores. Wolfe proposed that if the correlation coefficient of the two points of a computer score is significantly smaller than that of other computer scores, it indicates that the computer score is significantly different from other computer scores and is random. Correlation Pt Mea represents the two-column correlation coefficient for each computer score, with values between 0.36 and 0.76, with an average of 0.56, a standard deviation of 0.14, and a computer score of 7 The two-column correlation coefficient is below the standard deviation of 1.4 standard deviations, and there may be random effects. Other computer scores do not show obvious random effects, the halo effect evaluation at computer scoring is shown in Figure 5.

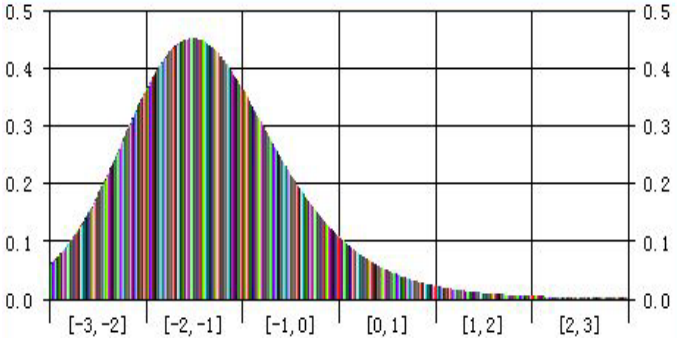

Figure 5. Halo effect on computer scoring

Halo effect refers to the fact that computer ratings cannot distinguish between different aspects of a player's ability to give similar scores, or use similar fractional computer scores. Computer scores from the statistical results of all aspects of the computer scores of the Table computer score 4 can be seen, the scores of the three scores of the three different aspects examined in the finals are the computer score 0.87 computer score, separation index for the computer A score of 1.49 on

computer scores indicates that the difficulty of the three aspects of the investigation can be roughly divided into the computer score 2 computer score levels, the separation reliability is very low, only computer score 0.43 computer scores, chi-square test computer scores (computer scores Chi-square = 3.5 computer score, computer score p = 0.17 computer score), indicating that the three levels do not have significant difference computer scores in terms of difficulty. Computer Scoring Thus, as a whole, computer scoring cannot effectively distinguish the three levels of players by computer scoring, and there is a significant halo effect computer score.

5. Conclusion

Based on the Rasch model's analysis of scoring behavior, we can see that computer scoring has a wide gap between the overall of the score. Among them, men are strict than women. Computer scoring individuals are basically consistent in terms of the degree of strictness of scores, but they are affected by experience, experience, personal feelings, etc. Individual computer scores such as No. 1 and No. 4 also show certain fluctuations. This means that computer scores need to be trained prior to scoring in order to maximize the objective and accurate assessment of students' writing skills.

Although this study uses a multi-level Rasch model to analyze the scoring behavior of writing, it can objectively reflect the information related to computer scoring. However, the sample selection is too simple, the number of samples is too small, and to some extent, it cannot fully reveal the score of computer scoring. Therefore, we need to further study, in order to better analyze the scoring behavior, and feedback the results to the rating teacher, so that rating teachers understand their scoring behavior, in order to improve in the future teaching and scoring work.

References

[1] Johansson S, Kottorp A, Lee K A, et al. Can the Fatigue Severity Scale 7-item version be used across different patient populations as a generic fatigue measure-a comparative study using a Rasch model approach[J]. Health and quality of life outcomes, 2014, 12(1): 24.

[2] Mahoney E R, Delaney C R. Regression Modeling System Using Activation Rating Values as Inputs to a Regression to Predict Healthcare Utilization and Cost and/or Changes Thereto: U.S. Patent Application 14/704,860[P]. 2015-11-5.

[3] Ghazzal, Mohamed Nawfal, et al. "Study of mesoporous CdS-quantum-dot-sensitized TiO2 films by using X-ray photoelectron spectroscopy and AFM." Beilstein journal of nanotechnology 5 (2014): 68.

[4] ElifKantarcıoglu, Carole Thomas, John O'Dwyer, and Barry O'Sullivan. "Benchmarking a high-stakes proficiency exam: the COPE linking project." Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual 33 (2010): 102.

[5] Aryadoust, Vahid. "Application of the fusion model to while-listening performance tests." Shiken: JALT Testing & Evaluation SIG Newsletter 15.2 (2011): 2-9.

[6] Baghaei, Purya, and Vahid Aryadoust. "Modeling local item dependence due to common test format with a multidimensional Rasch model." International Journal of Testing 15.1 (2015): 71-87.

[7] Varley, P. (2011).” Ecosophy and tourism: Rethinking a mountain resort”. Tourism Management, 32, 902-911

[8] Daud, Nor Shidrah Mat, Nuraihan Mat Daud, and Noor Lide Abu Kassim. "Second Language Writing Anxiety: Cause Or Effect?." Malaysian journal of ELT research 1.1 (2016): 19.