A Comparison of Two Scoring Methods of Chinese University Students' Writing in EFL Course Using MultiFacet Rasch Model

Meng Lyu[1,a]

[1] *Xi'an Jiaotong University*
Email: mariannelyu@stu.xjtu.edu.cn

ABSTRACT

Writing is one of the important productive skills for second language learners. Considering the inadequacy of the research on global scoring and descriptor-based checklist scoring in college English writing tests, this paper analyzes the scoring validity of the test at four dimensions: test-takers, rater, task, and scale and bias of rater-subject and rater-dimension. Based on the data of a college English writing test, the study will also employ the MultiFacet Rasch model (MFRM) to conduct a comparative analysis and bias analysis of the two scoring methods. The study showed that rater severity varied in global and descriptor-based ratings and had high internal consistency in descriptor-based scoring. This study can provide a reference for rater training and be used as a benchmark for deciding which scoring criteria to use in the L2 writing scoring for the EFL courses in China.

Keywords: *Writing assessment, L2 writing, MultiFacet Rasch Model, scoring validity*

1. INTRODUCTION

Writing is a key component of second language testing and one of the most common forms of testing. However, test validity is difficult to be guaranteed because of several factors. In addition to students' writing ability, the difficulty of the writing task, the assessment criteria and scales which the raters apply to evaluate the writing, and the subjective nature of the rating would affect not only the final score but also the corresponding validity of the writing tasks. Therefore, assessors need to collect different types of evidence and use it to demonstrate that the scoring methods are valid.

The divergence emerging from the definitions of validity and the standards to verify validity even worsens the situation. For example, construct validity, based on True Score Theory, examines the extent to which the test takers' performance matches the language proficiency theory envisioned by the test developer [1]. This index focuses on the factor structure within the test by using exploratory and validation factor methods [2,3]. Another type of rating validity is based on Item Response Theory, including rating stability and inter-rater and intra-rater consistency [4,5]. In comparison to the former, rating validity is a more realistic estimating procedure, which requires raters to track various sources of error in writing tests and produce accurate descriptions of the results. The MultiFacet Rasch Model (MFRM), which was developed and adapted by Linacre based on the Rasch model, has been frequently utilized in recent years for evaluating productive, subjective tests such as speaking [6,7,8] and writing [9,10]. The computational logic of the model is to unify the estimates of each dimension in the same measurement panel so that they can be calculated linearly and be compared cross-dimensionally. As a result, the accuracy of the test scores depends on the difference among the ability of the test takers, the difficulty of the task, the severity of the rater, and the rating scale [11].

As a small-scale test, English writing assignments in college courses have gradually received more attention, while there have been a few studies on the validity of the rating scheme of this task. In general, a few studies have investigated the application of the Rasch model to measure writing tasks in college English courses in China [12].

To summarize, this paper aims to delve into the scoring validity of the propositional essay writing exam. The following research questions will be addressed in this study:

1) Whether the different scoring methods affect the validity of scoring results in the English writing test?

2) Is the scoring behavior of the scorers affected by the two different scoring methods, and is their scoring behavior consistent and stable?

In order to provide a comprehensive account of these research questions, this paper will analyze the phenomenon in aspects of subject differentiation, rater severity, and rating scale.

2. METHOD

2.1 Participants and Writing Prompt

32 first-year students were recruited as the research participants, and their final writing scores in the college English writing courses were selected as the data source. In the writing task, the research subjects were required to write a 120-word essay within 30 minutes on the difficult level equivalent to Chinese College English Test Band 4.

2.2 Scoring Criteria and Data Collection

The essays were marked by 10 raters who were all graduate students in applied linguistics. The global scoring scale applied in the study was divided into three dimensions, which examined the quality of writing in terms of content, organization, and language use, respectively. The descriptor-based checklist was composed of 30 descriptors in 5 dimensions, namely, mechanics, vocabulary, content, organization, and language use.

Both scales contain five levels, ranging from1 to 5, with a 1-point scoring interval. This is intended to eliminate over-centered scoring. To maintain impartiality in grading, the information of the test takers was excluded from the essay, and all raters received rating training prior to scoring. Also, all raters employed two different scoring methods to assess the same writing.

3. RESULTS

3.1 Comparison analysis of the global and the descriptor-based scoring

Table 1 and Table 2 demonstrate the overall situation of the global scoring model and the descriptor-based scoring model, respectively.

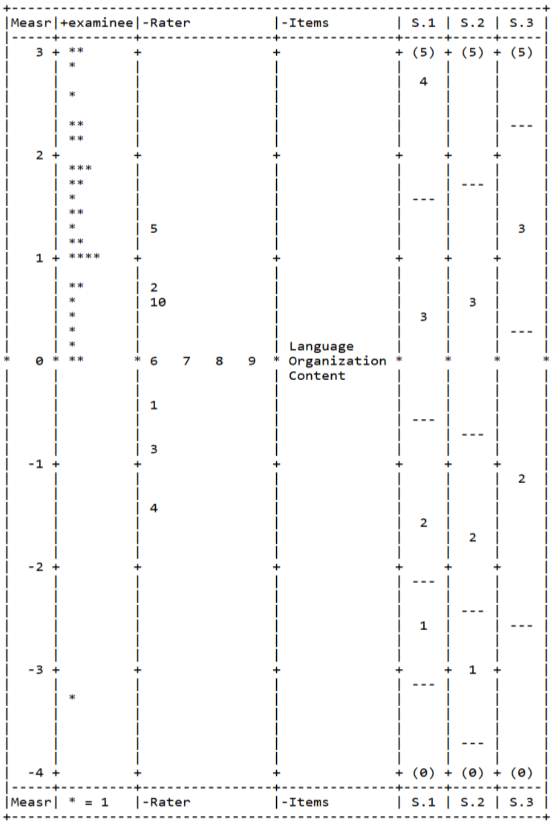
In Table I and Table II, the span of ability distribution (column 2) is substantial (approximately 3.2 logits), indicating that the global scoring can clearly identify and distinguish subjects' writing abilities.

The severity of raters (column 3) in Table I and II is basically normally distributed, with rater 5 being the strictest and rater 4 the loosest in global scoring. There is a high level of inter-rater agreement in both rating methods.

Concerning the level of the rating scale (column 4), Table 1 demonstrates that the subjects had the most difficulty in obtaining high scores in language use as the raters are stricter in assessing this item (about 0.2 logits), while in descriptor-based scoring, vocabulary (approximately 0.3 logits) is the most difficult.

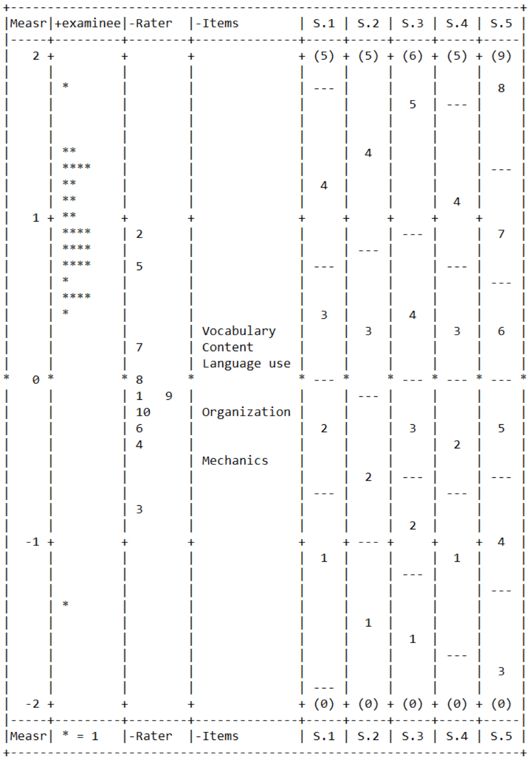
The boundaries among each of the five items are relatively evident in both methods; For organization in global scoring, band 4 is not easily distinguishable from the proximity score; whereas for language use in descriptor-based scoring, band 1 and band 4 are difficult to differentiate from the proximity score.

Table 1. MFRM analysis of holistic scoring



```
+-----------------------------------------------------------------+
|Measr|+examinee|-Rater          |-Items        | S.1 | S.2 | S.3 |
|-----+---------+----------------+--------------+-----+-----+-----|
|   3 + **      +                +              + (5) + (5) + (5) |
|     | *       |                |              |     |     |     |
|     |         |                |              |  4  |     |     |
|     | *       |                |              |     |     |     |
|     |         |                |              |     |     |     |
|     | **      |                |              |     |     | --- |
|     | **      |                |              |     |     |     |
|   2 +         +                +              +     +     +     |
|     | ***     |                |              |     |     |     |
|     | **      |                |              |     | --- |     |
|     | *       |                |              | --- |     |     |
|     | **      |                |              |     |     |     |
|     | *       | 5              |              |     |     |  3  |
|     | **      |                |              |     |     |     |
|   1 + ****    +                +              +     +     +     |
|     |         |                |              |     |     |     |
|     | **      | 2              |              |     |     |     |
|     | *       | 10             |              |     |  3  |     |
|     | *       |                |              |  3  |     |     |
|     | *       |                |              |     |     | --- |
|     | *       |                | Language     |     |     |     |
*   0 * **      * 6   7   8   9  * Organization *     *     *     *
|     |         |                | Content      |     |     |     |
|     |         |                |              |     |     |     |
|     |         | 1              |              |     |     |     |
|     |         |                |              | --- |     |     |
|     |         |                |              |     | --- |     |
|     |         | 3              |              |     |     |     |
|  -1 +         +                +              +     +     +     |
|     |         |                |              |     |     |  2  |
|     |         |                |              |     |     |     |
|     |         | 4              |              |     |     |     |
|     |         |                |              |  2  |     |     |
|     |         |                |              |     |  2  |     |
|     |         |                |              |     |     |     |
|  -2 +         +                +              +     +     +     |
|     |         |                |              | --- |     |     |
|     |         |                |              |     |     |     |
|     |         |                |              |     | --- |     |
|     |         |                |              |  1  |     | --- |
|     |         |                |              |     |     |     |
|     |         |                |              |     |     |     |
|  -3 +         +                +              +     +  1  +     |
|     |         |                |              | --- |     |     |
|     | *       |                |              |     |     |     |
|     |         |                |              |     |     |     |
|     |         |                |              |     |     |     |
|     |         |                |              |     | --- |     |
|     |         |                |              |     |     |     |
|  -4 +         +                +              + (0) + (0) + (0) |
|-----+---------+----------------+--------------+-----+-----+-----|
|Measr| * = 1   |-Rater          |-Items        | S.1 | S.2 | S.3 |
+-----------------------------------------------------------------+
```

Table 2. MFRM analysis of descriptor-based scoring



```
+---------------------------------------------------------------------+
|Measr|+examinee|-Rater  |-Items        | S.1 | S.2 | S.3 | S.4 | S.5 |
|-----+---------+--------+--------------+-----+-----+-----+-----+-----|
|   2 +         +        +              + (5) + (5) + (6) + (5) + (9) |
|     |         |        |              |     |     |     |     |     |
|     | *       |        |              | --- |     |     |     |  8  |
|     |         |        |              |     |     |  5  | --- |     |
|     |         |        |              |     |     |     |     |     |
|     |         |        |              |     |     |     |     |     |
|     | **      |        |              |     |  4  |     |     |     |
|     | ****    |        |              |     |     |     |     | --- |
|     | **      |        |              |  4  |     |     |     |     |
|     | **      |        |              |     |     |     |  4  |     |
|   1 + **      +        +              +     +     +     +     +     |
|     | ****    | 2      |              |     |     | --- |     |  7  |
|     | ****    |        |              |     | --- |     |     |     |
|     | ****    | 5      |              | --- |     |     | --- |     |
|     | *       |        |              |     |     |     |     | --- |
|     | ****    |        |              |     |     |     |     |     |
|     | *       |        |              |  3  |     |  4  |     |     |
|     |         |        | Vocabulary   |     |  3  |     |  3  |  6  |
|     |         | 7      | Content      |     |     |     |     |     |
|     |         |        | Language use |     |     |     |     |     |
*   0 *         * 8      *              * --- *     * --- * --- * --- *
|     |         | 1   9  |              |     | --- |     |     |     |
|     |         | 10     | Organization |     |     |     |     |     |
|     |         | 6      |              |  2  |     |  3  |     |  5  |
|     |         | 4      |              |     |     |     |  2  |     |
|     |         |        | Mechanics    |     |     |     |     |     |
|     |         |        |              |     |  2  | --- |     | --- |
|     |         |        |              | --- |     |     | --- |     |
|     |         | 3      |              |     |     |     |     |     |
|     |         |        |              |     |     |  2  |     |     |
|  -1 +         +        +              +     + --- +     +     +  4  |
|     |         |        |              |  1  |     |     |  1  |     |
|     |         |        |              |     |     | --- |     |     |
|     |         |        |              |     |     |     |     | --- |
|     | *       |        |              |     |     |     |     |     |
|     |         |        |              |     |  1  |     |     |     |
|     |         |        |              |     |     |  1  |     |     |
|     |         |        |              |     |     |     | --- |     |
|     |         |        |              |     |     |     |     |  3  |
|     |         |        |              | --- |     |     |     |     |
|  -2 +         +        +              + (0) + (0) + (0) + (0) + (0) |
|-----+---------+--------+--------------+-----+-----+-----+-----+-----|
|Measr| * = 1   |-Rater  |-Items        | S.1 | S.2 | S.3 | S.4 | S.5 |
+---------------------------------------------------------------------+
```

3.2 Analysis of the rater effects

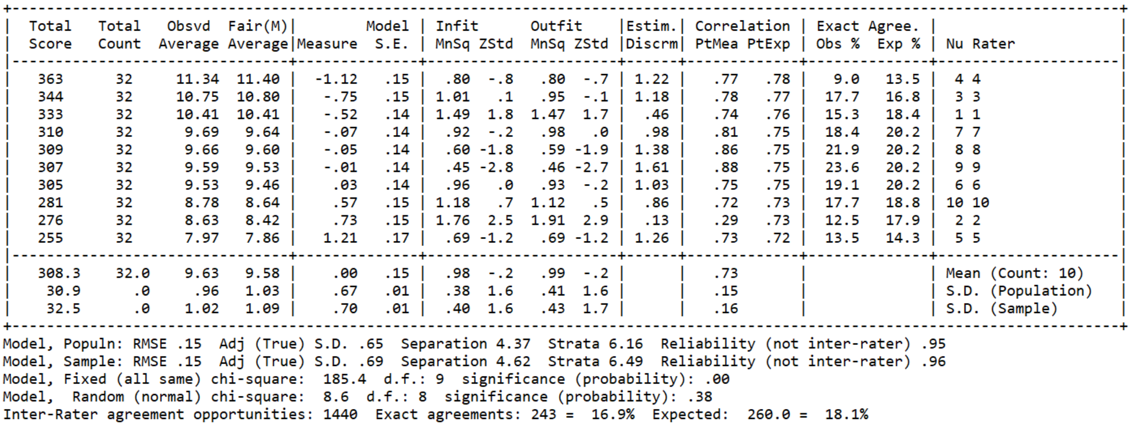
As seen in Table III and Table IV, the raters are ranked in order of severity level from -1.12 to 1.21 logits, with a range of 3.33 logits. Rater 5 is the most severe, followed by rater 2, 10, 6, 9, 8, 7, 1, and 3, with rater 4 being the most lenient (roughly -0.21 logits). In Table IV, raters range from -0.71 to 0.97 logits, with a range of 1.68 logits. The most severe is rater 5, while the least severe is rater 2. The information in the table appropriately depicts the severity disparities among the raters.

At the bottom of Table III and IV, the separation coefficient, reliability, and chi-square values are presented, which indicate the actual degree of difference in severity among the raters. The separation coefficients in Table 3 are 4.37 and 4.62 (greater than 2), revealing that the 10 raters could be divided into at least four severity categories with considerable disparities in severity amongst raters. However, the chi-square value of 185.4 for d.f. 9 is significant at a significance value of 0.00 and a reliability of 0.95.

There are two types of rater reliability, including inter-rater reliability referring to the agreement among the raters and intra-rater reliability referring to self-consistency within one rater. In Table IV, with the exception of rater 2, all remaining nine raters reached a high level of internal consistency. Infit MnSq values imply that rater variation is likely to be lower than the model predicts as rater ratings tend to be clustered, centered, and overfitted.

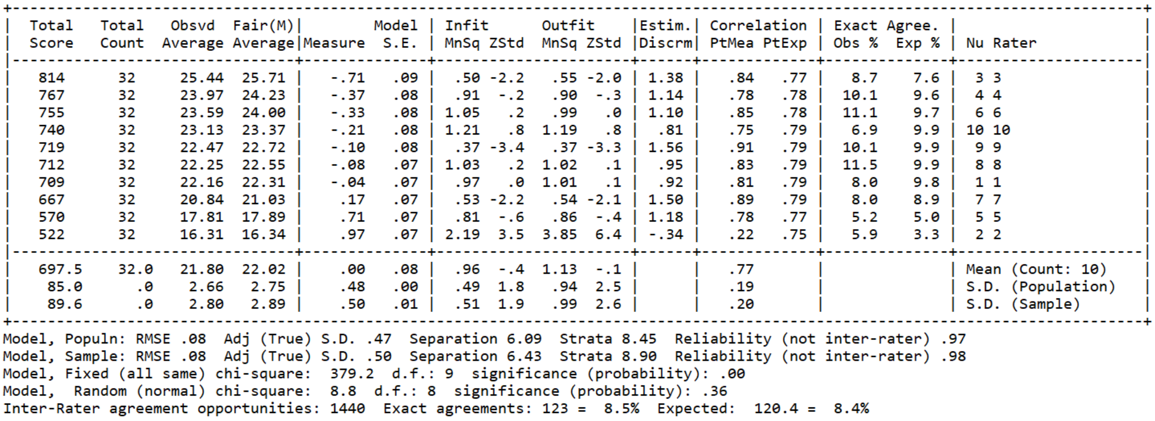
In Table III, only rater 9 and rater 2 did not fall within the interval of 0.5 to 1.5 in ratings. Rater 9 is less than 0.5, indicating that the rating is convergent, while rater 2 is greater than 1.5, indicating that the rater has poor internal consistency. This may be due to the rater's inexperience with scoring and the insufficient awareness of the scoring criteria. Also, the scoring process was divided into two sessions, which may result in changes in their criteria. However, the majority of the raters were reliable with high degrees of self-consistency.

Table 3. Rater measurement of global scoring



```
+-------------------------------------------------------------------------------------------------------------------------------+
|  Total    Total    Obsvd  Fair(M)|          Model | Infit       Outfit     |Estim.| Correlation | Exact Agree. |                      |
|  Score    Count  Average Average|Measure   S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %   Exp % | Nu Rater             |
|---------------------------------+----------------+----------------------+------+-------------+--------------+----------------------|
|   363       32     11.34   11.40 |   -1.12    .15 |  .80   -.8    .80   -.7 | 1.22 |   .77   .78 |   9.0   13.5 |  4 4                 |
|   344       32     10.75   10.80 |    -.75    .15 | 1.01    .1    .95   -.1 | 1.18 |   .78   .77 |  17.7   16.8 |  3 3                 |
|   333       32     10.41   10.41 |    -.52    .14 | 1.49   1.8   1.47   1.7 |  .46 |   .74   .76 |  15.3   18.4 |  1 1                 |
|   310       32      9.69    9.64 |    -.07    .14 |  .92   -.2    .98    .0 |  .98 |   .81   .75 |  18.4   20.2 |  7 7                 |
|   309       32      9.66    9.60 |    -.05    .14 |  .60  -1.8    .59  -1.9 | 1.38 |   .86   .75 |  21.9   20.2 |  8 8                 |
|   307       32      9.59    9.53 |    -.01    .14 |  .45  -2.8    .46  -2.7 | 1.61 |   .88   .75 |  23.6   20.2 |  9 9                 |
|   305       32      9.53    9.46 |     .03    .14 |  .96    .0    .93   -.2 | 1.03 |   .75   .75 |  19.1   20.2 |  6 6                 |
|   281       32      8.78    8.64 |     .57    .15 | 1.18    .7   1.12    .5 |  .86 |   .72   .73 |  17.7   18.8 | 10 10                |
|   276       32      8.63    8.42 |     .73    .15 | 1.76   2.5   1.91   2.9 |  .13 |   .29   .73 |  12.5   17.9 |  2 2                 |
|   255       32      7.97    7.86 |    1.21    .17 |  .69  -1.2    .69  -1.2 | 1.26 |   .73   .72 |  13.5   14.3 |  5 5                 |
|---------------------------------+----------------+----------------------+------+-------------+--------------+----------------------|
|   308.3    32.0     9.63    9.58 |     .00    .15 |  .98   -.2    .99   -.2 |      |   .73       |              | Mean (Count: 10)     |
|    30.9      .0      .96    1.03 |     .67    .01 |  .38   1.6    .41   1.6 |      |   .15       |              | S.D. (Population)    |
|    32.5      .0     1.02    1.09 |     .70    .01 |  .40   1.6    .43   1.7 |      |   .16       |              | S.D. (Sample)        |
+-------------------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .15  Adj (True) S.D. .65  Separation 4.37  Strata 6.16  Reliability (not inter-rater) .95
Model, Sample: RMSE .15  Adj (True) S.D. .69  Separation 4.62  Strata 6.49  Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square:  185.4  d.f.: 9  significance (probability): .00
Model,  Random (normal) chi-square:  8.6  d.f.: 8  significance (probability): .38
Inter-Rater agreement opportunities: 1440  Exact agreements: 243 =  16.9%  Expected:  260.0 =  18.1%
```

Table 4. Rater measurement of descriptor-based scoring

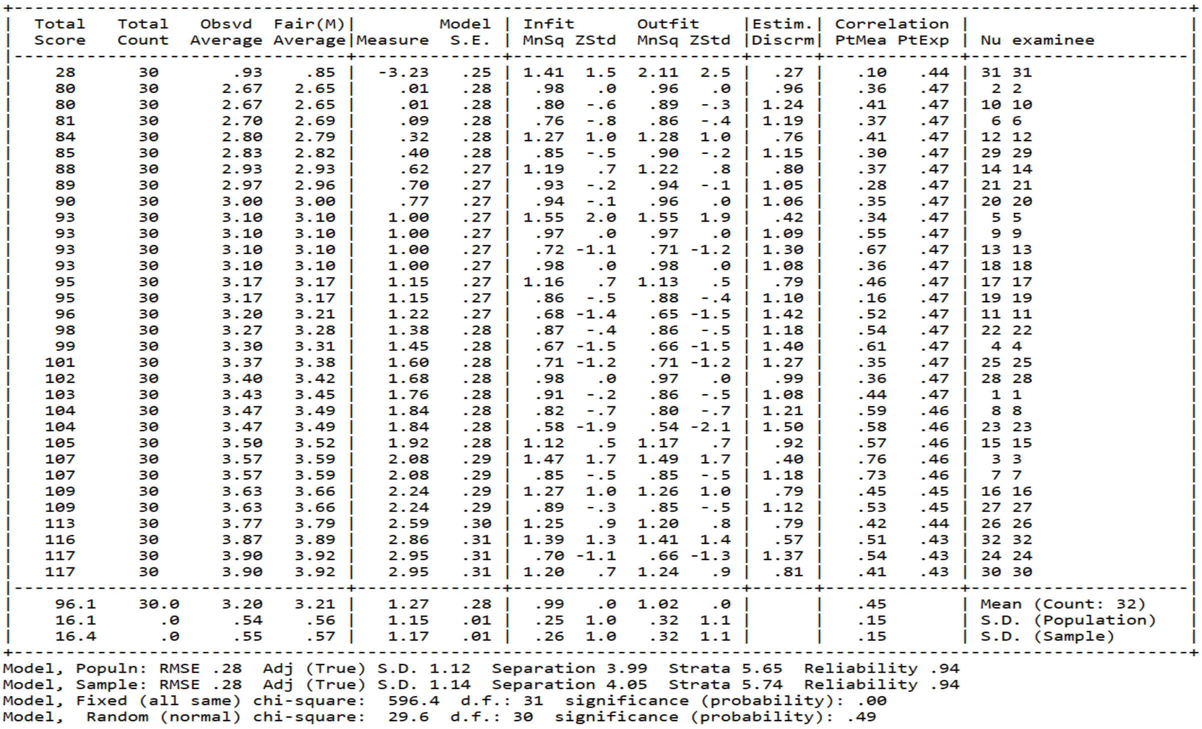


```
+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation | Exact Agree. |                   |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %  Exp % | Nu Rater          |
|--------------------------------+--------------+-----------------------+------+-------------+--------------+-------------------|
|   814      32    25.44  25.71  |   -.71   .09 |  .50 -2.2   .55 -2.0 | 1.38 |   .84   .77 |   8.7    7.6 |  3 3              |
|   767      32    23.97  24.23  |   -.37   .08 |  .91  -.2   .90  -.3 | 1.14 |   .78   .78 |  10.1    9.6 |  4 4              |
|   755      32    23.59  24.00  |   -.33   .08 | 1.05   .2   .99   .0 | 1.10 |   .85   .78 |  11.1    9.7 |  6 6              |
|   740      32    23.13  23.37  |   -.21   .08 | 1.21   .8  1.19   .8 |  .81 |   .75   .79 |   6.9    9.9 | 10 10             |
|   719      32    22.47  22.72  |   -.10   .08 |  .37 -3.4   .37 -3.3 | 1.56 |   .91   .79 |  10.1    9.9 |  9 9              |
|   712      32    22.25  22.55  |   -.08   .07 | 1.03   .2  1.02   .1 |  .95 |   .83   .79 |  11.5    9.9 |  8 8              |
|   709      32    22.16  22.31  |   -.04   .07 |  .97   .0  1.01   .1 |  .92 |   .81   .79 |   8.0    9.8 |  1 1              |
|   667      32    20.84  21.03  |    .17   .07 |  .53 -2.2   .54 -2.1 | 1.50 |   .89   .79 |   8.0    8.9 |  7 7              |
|   570      32    17.81  17.89  |    .71   .07 |  .81  -.6   .86  -.4 | 1.18 |   .78   .77 |   5.2    5.0 |  5 5              |
|   522      32    16.31  16.34  |    .97   .07 | 2.19  3.5  3.85  6.4 | -.34 |   .22   .75 |   5.9    3.3 |  2 2              |
|--------------------------------+--------------+-----------------------+------+-------------+--------------+-------------------|
|   697.5    32.0  21.80  22.02  |    .00   .08 |  .96  -.4  1.13  -.1 |      |   .77       |              | Mean (Count: 10)  |
|    85.0      .0   2.66   2.75  |    .48   .00 |  .49  1.8   .94  2.5 |      |   .19       |              | S.D. (Population) |
|    89.6      .0   2.80   2.89  |    .50   .01 |  .51  1.9   .99  2.6 |      |   .20       |              | S.D. (Sample)     |
+-------------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .08  Adj (True) S.D. .47  Separation 6.09  Strata 8.45  Reliability (not inter-rater) .97
Model, Sample: RMSE .08  Adj (True) S.D. .50  Separation 6.43  Strata 8.90  Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square:  379.2  d.f.: 9  significance (probability): .00
Model,  Random (normal) chi-square:  8.8  d.f.: 8  significance (probability): .36
Inter-Rater agreement opportunities: 1440  Exact agreements: 123 =  8.5%  Expected:  120.4 =  8.4%
```

3.3 Analysis of the examinee effects

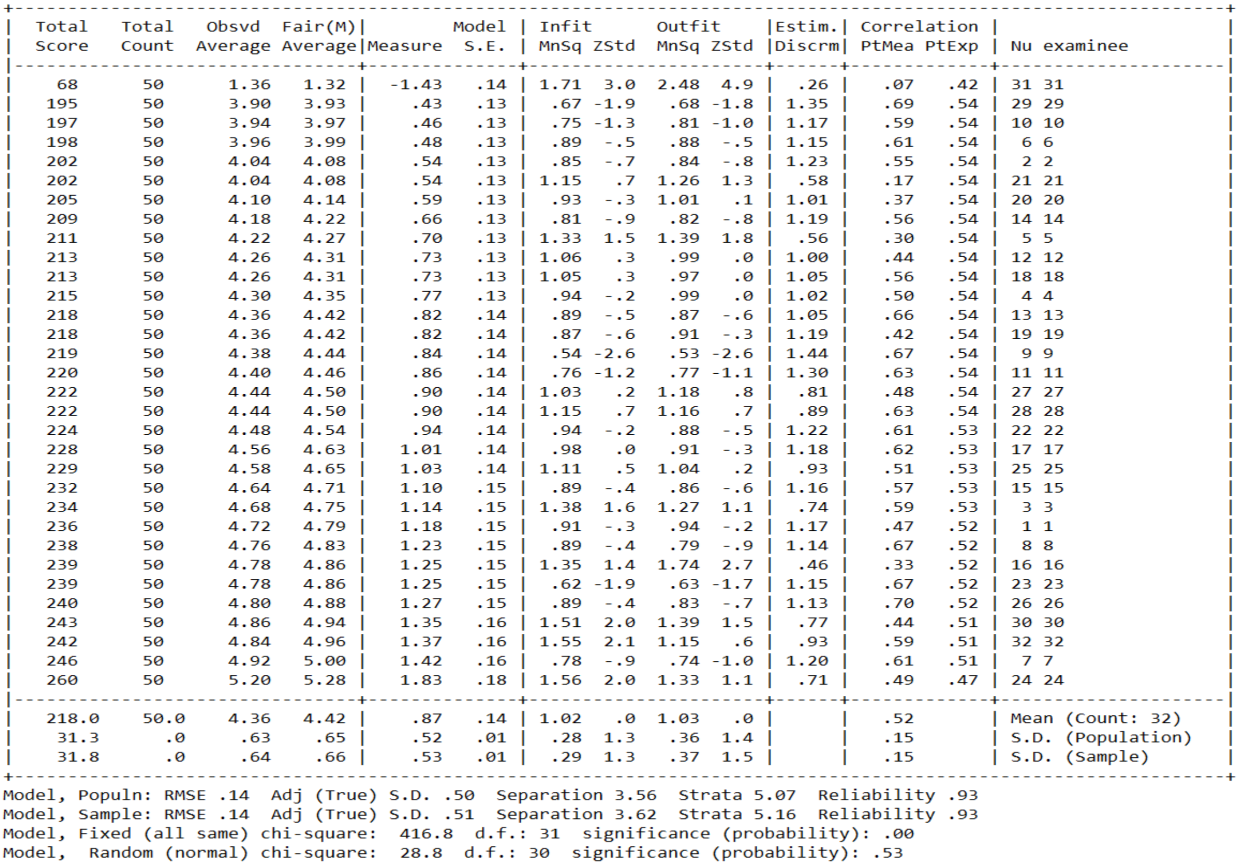
With regards to the subject-level analysis in Table V, the separation coefficient was 3.99, and the separation reliability was 0.94, suggesting those established marking criteria can distinguish the participants' writing skills considerably. With all items being fittingly less than 1.5, the raters were proven to be consistent in evaluating these essays; however, the overfitting items of rater 5 demonstrated a convergence or halo effect when the rater reviewed the essays. Also, a reliable model fit can be observed with a merely 0.28 error value. However, because of the absence of overfitting figures, Table VI displayed no neutralization or halo effect when the scorers reviewed this essay. The error of the model fit was 0.14, which can also be considered a trustworthy model fit.

Table 5. Measurement report of the global scoring



```
+-----------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation |                   |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu examinee       |
|--------------------------------+--------------+-----------------------+------+-------------+-------------------|
|    28     30       .93    .85  |  -3.23   .25 | 1.41  1.5  2.11  2.5 |  .27 |   .10   .44 | 31 31             |
|    80     30      2.67   2.65  |    .01   .28 |  .98   .0   .96   .0 |  .96 |   .36   .47 |  2 2              |
|    80     30      2.67   2.65  |    .01   .28 |  .80  -.6   .89  -.3 | 1.24 |   .41   .47 | 10 10             |
|    81     30      2.70   2.69  |    .09   .28 |  .76  -.8   .86  -.4 | 1.19 |   .37   .47 |  6 6              |
|    84     30      2.80   2.79  |    .32   .28 | 1.27  1.0  1.28  1.0 |  .76 |   .41   .47 | 12 12             |
|    85     30      2.83   2.82  |    .40   .28 |  .85  -.5   .90  -.2 | 1.15 |   .30   .47 | 29 29             |
|    88     30      2.93   2.93  |    .62   .27 | 1.19   .7  1.22   .8 |  .80 |   .37   .47 | 14 14             |
|    89     30      2.97   2.96  |    .70   .27 |  .93  -.2   .94  -.1 | 1.05 |   .28   .47 | 21 21             |
|    90     30      3.00   3.00  |    .77   .27 |  .94  -.1   .96   .0 | 1.06 |   .35   .47 | 20 20             |
|    93     30      3.10   3.10  |   1.00   .27 | 1.55  2.0  1.55  1.9 |  .42 |   .34   .47 |  5 5              |
|    93     30      3.10   3.10  |   1.00   .27 |  .97   .0   .97   .0 | 1.09 |   .55   .47 |  9 9              |
|    93     30      3.10   3.10  |   1.00   .27 |  .72 -1.1   .71 -1.2 | 1.30 |   .67   .47 | 13 13             |
|    93     30      3.10   3.10  |   1.00   .27 |  .98   .0   .98   .0 | 1.08 |   .36   .47 | 18 18             |
|    95     30      3.17   3.17  |   1.15   .27 | 1.16   .7  1.13   .5 |  .79 |   .46   .47 | 17 17             |
|    95     30      3.17   3.17  |   1.15   .27 |  .86  -.5   .88  -.4 | 1.10 |   .16   .47 | 19 19             |
|    96     30      3.20   3.21  |   1.22   .27 |  .68 -1.4   .65 -1.5 | 1.42 |   .52   .47 | 11 11             |
|    98     30      3.27   3.28  |   1.38   .28 |  .87  -.4   .86  -.5 | 1.18 |   .54   .47 | 22 22             |
|    99     30      3.30   3.31  |   1.45   .28 |  .67 -1.5   .66 -1.5 | 1.40 |   .61   .47 |  4 4              |
|   101     30      3.37   3.38  |   1.60   .28 |  .71 -1.2   .71 -1.2 | 1.27 |   .35   .47 | 25 25             |
|   102     30      3.40   3.42  |   1.68   .28 |  .98   .0   .97   .0 |  .99 |   .36   .47 | 28 28             |
|   103     30      3.43   3.45  |   1.76   .28 |  .91  -.2   .86  -.5 | 1.08 |   .44   .47 |  1 1              |
|   104     30      3.47   3.49  |   1.84   .28 |  .82  -.7   .80  -.7 | 1.21 |   .59   .46 |  8 8              |
|   104     30      3.47   3.49  |   1.84   .28 |  .58 -1.9   .54 -2.1 | 1.50 |   .58   .46 | 23 23             |
|   105     30      3.50   3.52  |   1.92   .28 | 1.12   .5  1.17   .7 |  .92 |   .57   .46 | 15 15             |
|   107     30      3.57   3.59  |   2.08   .29 | 1.47  1.7  1.49  1.7 |  .40 |   .76   .46 |  3 3              |
|   107     30      3.57   3.59  |   2.08   .29 |  .85  -.5   .85  -.5 | 1.18 |   .73   .46 |  7 7              |
|   109     30      3.63   3.66  |   2.24   .29 | 1.27  1.0  1.26  1.0 |  .79 |   .45   .45 | 16 16             |
|   109     30      3.63   3.66  |   2.24   .29 |  .89  -.3   .85  -.5 | 1.12 |   .53   .45 | 27 27             |
|   113     30      3.77   3.79  |   2.59   .30 | 1.25   .9  1.20   .8 |  .79 |   .42   .44 | 26 26             |
|   116     30      3.87   3.89  |   2.86   .31 | 1.39  1.3  1.41  1.4 |  .57 |   .51   .43 | 32 32             |
|   117     30      3.90   3.92  |   2.95   .31 |  .70 -1.1   .66 -1.3 | 1.37 |   .54   .43 | 24 24             |
|   117     30      3.90   3.92  |   2.95   .31 | 1.20   .7  1.24   .9 |  .81 |   .41   .43 | 30 30             |
|--------------------------------+--------------+-----------------------+------+-------------+-------------------|
|    96.1   30.0    3.20   3.21  |   1.27   .28 |  .99   .0  1.02   .0 |      |   .45       | Mean (Count: 32)  |
|    16.1     .0     .54    .56  |   1.15   .01 |  .25  1.0   .32  1.1 |      |   .15       | S.D. (Population) |
|    16.4     .0     .55    .57  |   1.17   .01 |  .26  1.0   .32  1.1 |      |   .15       | S.D. (Sample)     |
+-----------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .28  Adj (True) S.D. 1.12  Separation 3.99  Strata 5.65  Reliability .94
Model, Sample: RMSE .28  Adj (True) S.D. 1.14  Separation 4.05  Strata 5.74  Reliability .94
Model, Fixed (all same) chi-square:  596.4  d.f.: 31  significance (probability): .00
Model,  Random (normal) chi-square:  29.6  d.f.: 30  significance (probability): .49
```

Table 6. Measurement report of the global scoring



```
+-----------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit    |Estim.| Correlation |                        |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu examinee            |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|    68      50      1.36   1.32 |  -1.43   .14 | 1.71  3.0  2.48  4.9 |  .26 |   .07   .42 | 31 31                  |
|   195      50      3.90   3.93 |    .43   .13 |  .67 -1.9   .68 -1.8 | 1.35 |   .69   .54 | 29 29                  |
|   197      50      3.94   3.97 |    .46   .13 |  .75 -1.3   .81 -1.0 | 1.17 |   .59   .54 | 10 10                  |
|   198      50      3.96   3.99 |    .48   .13 |  .89  -.5   .88  -.5 | 1.15 |   .61   .54 |  6 6                   |
|   202      50      4.04   4.08 |    .54   .13 |  .85  -.7   .84  -.8 | 1.23 |   .55   .54 |  2 2                   |
|   202      50      4.04   4.08 |    .54   .13 | 1.15   .7  1.26  1.3 |  .58 |   .17   .54 | 21 21                  |
|   205      50      4.10   4.14 |    .59   .13 |  .93  -.3  1.01   .1 | 1.01 |   .37   .54 | 20 20                  |
|   209      50      4.18   4.22 |    .66   .13 |  .81  -.9   .82  -.8 | 1.19 |   .56   .54 | 14 14                  |
|   211      50      4.22   4.27 |    .70   .13 | 1.33  1.5  1.39  1.8 |  .56 |   .30   .54 |  5 5                   |
|   213      50      4.26   4.31 |    .73   .13 | 1.06   .3   .99   .0 | 1.00 |   .44   .54 | 12 12                  |
|   213      50      4.26   4.31 |    .73   .13 | 1.05   .3   .97   .0 | 1.05 |   .56   .54 | 18 18                  |
|   215      50      4.30   4.35 |    .77   .13 |  .94  -.2   .99   .0 | 1.02 |   .50   .54 |  4 4                   |
|   218      50      4.36   4.42 |    .82   .14 |  .89  -.5   .87  -.6 | 1.05 |   .66   .54 | 13 13                  |
|   218      50      4.36   4.42 |    .82   .14 |  .87  -.6   .91  -.3 | 1.19 |   .42   .54 | 19 19                  |
|   219      50      4.38   4.44 |    .84   .14 |  .54 -2.6   .53 -2.6 | 1.44 |   .67   .54 |  9 9                   |
|   220      50      4.40   4.46 |    .86   .14 |  .76 -1.2   .77 -1.1 | 1.30 |   .63   .54 | 11 11                  |
|   222      50      4.44   4.50 |    .90   .14 | 1.03   .2  1.18   .8 |  .81 |   .48   .54 | 27 27                  |
|   222      50      4.44   4.50 |    .90   .14 | 1.15   .7  1.16   .7 |  .89 |   .63   .54 | 28 28                  |
|   224      50      4.48   4.54 |    .94   .14 |  .94  -.2   .88  -.5 | 1.22 |   .61   .53 | 22 22                  |
|   228      50      4.56   4.63 |   1.01   .14 |  .98   .0   .91  -.3 | 1.18 |   .62   .53 | 17 17                  |
|   229      50      4.58   4.65 |   1.03   .14 | 1.11   .5  1.04   .2 |  .93 |   .51   .53 | 25 25                  |
|   232      50      4.64   4.71 |   1.10   .15 |  .89  -.4   .86  -.6 | 1.16 |   .57   .53 | 15 15                  |
|   234      50      4.68   4.75 |   1.14   .15 | 1.38  1.6  1.27  1.1 |  .74 |   .59   .53 |  3 3                   |
|   236      50      4.72   4.79 |   1.18   .15 |  .91  -.3   .94  -.2 | 1.17 |   .47   .52 |  1 1                   |
|   238      50      4.76   4.83 |   1.23   .15 |  .89  -.4   .79  -.9 | 1.14 |   .67   .52 |  8 8                   |
|   239      50      4.78   4.86 |   1.25   .15 | 1.35  1.4  1.74  2.7 |  .46 |   .33   .52 | 16 16                  |
|   239      50      4.78   4.86 |   1.25   .15 |  .62 -1.9   .63 -1.7 | 1.15 |   .67   .52 | 23 23                  |
|   240      50      4.80   4.88 |   1.27   .15 |  .89  -.4   .83  -.7 | 1.13 |   .70   .52 | 26 26                  |
|   243      50      4.86   4.94 |   1.35   .16 | 1.51  2.0  1.39  1.5 |  .77 |   .44   .51 | 30 30                  |
|   242      50      4.84   4.96 |   1.37   .16 | 1.55  2.1  1.15   .6 |  .93 |   .59   .51 | 32 32                  |
|   246      50      4.92   5.00 |   1.42   .16 |  .78  -.9   .74 -1.0 | 1.20 |   .61   .51 |  7 7                   |
|   260      50      5.20   5.28 |   1.83   .18 | 1.56  2.0  1.33  1.1 |  .71 |   .49   .47 | 24 24                  |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|   218.0    50.0    4.36   4.42 |    .87   .14 | 1.02   .0  1.03   .0 |      |   .52       | Mean (Count: 32)       |
|    31.3      .0     .63    .65 |    .52   .01 |  .28  1.3   .36  1.4 |      |   .15       | S.D. (Population)      |
|    31.8      .0     .64    .66 |    .53   .01 |  .29  1.3   .37  1.5 |      |   .15       | S.D. (Sample)          |
+-----------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .14  Adj (True) S.D. .50  Separation 3.56  Strata 5.07  Reliability .93
Model, Sample: RMSE .14  Adj (True) S.D. .51  Separation 3.62  Strata 5.16  Reliability .93
Model, Fixed (all same) chi-square:  416.8  d.f.: 31  significance (probability): .00
Model,  Random (normal) chi-square:  28.8  d.f.: 30  significance (probability): .53
```

3.4 Analysis of the two rating scales

The model analysis provides probability plots of the scores for the different dimensions of the two scoring methods.

In global scoring, in content rating, the peaks of band 1 are not obvious, while organization and language use have distinct coverage regions and separate peaks, suggesting a scale of 1 to 5, it is difficult for the raters to separate band 1 from band 2, whereas the other ratings are easier to be provided.

In descriptor-based scoring, in general, the probability curves for the five dimensions exhibit significant peaks for scores from 1 to 5, indicating that the raters are able to capture the differences in these scores. To be more specific, in mechanics and language use, raters are able to grasp the differences between these bands well. In vocabulary, it is difficult for the raters to separate the band 2 from the neighbor band 4, while the other scores are easier to judge. In content and organization, all scores have sharp peaks except for the band 2, which is not obvious enough.

3.5 Bias Analysis

3.5.1 Bias in the rater-subject interaction

There were 320 experimental bias items, and Table 7 shows that there were 21 significant biases (11 of which were too severe and 10 of which were too generous), accounting for 6.5% , slightly above the acceptable range of 5%. The raters with significant bias were rater 1 (2 times too severe and 4 times too generous) and rater 2 (3 times too severe and 3 times too generous). These raters need focused training. The top raters were rater 3, 4, 5, 8, and 9, who scored very consistently with no bias.

In addition, according to the results of the rater-subject bias analysis, the rater deviated the most from subject 31 (3 serious bias), and the scores given by the rater ranged from 0 to 7. The topic of the essay is *The challenge of living in a big city*, but subject 31's essay only covers telephone calls, which is a deviation from the topic, and is probably a copy from the reading comprehension text. It can be found that the content, organization and language use of the essay are relatively neat, except for the main idea which does not meet the requirements of the essay. The relatively large scoring errors may be due to the influence of each dimension.

Table 7. Rater-subject bias

Rater	Severity	Test-taker	Difficulty	Bias
2	0.75	31	-3.23	2.97
2	0.75	14	0.62	2.64
6	0.07	28	1.68	2.45
1	-0.43	3	2.08	2.4
6	0.07	5	1	2.24
3	-0.88	9	1	2.21
5	1.31	31	-3.23	2.21
7	-0.05	8	1.84	2.21
10	0.64	26	2.59	2.15
1	-0.43	26	2.59	2.09
7	-0.05	12	0.32	2.09
7	-0.05	31	-3.23	-1.52
4	-1.39	21	0.7	-2.33
1	-0.43	9	1	-2.43
1	-0.43	18	1	-2.43
6	0.07	30	2.95	-2.43
10	0.64	3	2.08	-2.44
1	1.15	19	1.15	-2.59
1	-0.43	15	1.92	-2.62
8	-0.03	5	1	-3.11
2	0.75	32	2.86	-3.17

3.5.2 Bias in the rater-dimension interaction

There were 30 experimental bias items, among which 5 significant biases, accounting for 16.6%, were above the acceptable range. Table 8 shows the results of the bias analysis between the scoring dimensions and rater. There were five significant biases between the raters on three dimensions, and the bias mainly lied in language and content, of which three were too generous and two were too severe. Raters 2 and 7 showed more biased on the scoring, indicating that both raters lacked a deep understanding of the scoring criteria. This information suggests that some of the raters had an unstable grasp of severity in evaluating writings.

Table 8. Rater-dimensions Bias

Rater	Severity	Dimensions	Difficulty	Bias
2	0.75	Language	0.13	3.28
7	-0.05	Organization	0	2.1
2	0.75	Content	-0.13	-3.07
10	0.64	Language	0.13	-2.11
7	-0.05	Content	-0.13	-2.66

3.5.3 Comparison of bias analysis of two rating methods

A longitudinal comparison of rater performance revealed that rater 2 and 7 were prone to bias in both overall and checklist-based scale, and that rater 2 was overly severe in both rating methods, and Rater 7 was more severe when using checklist-based scale than overall rating. Rater 1, who had the most severe bias in overall scoring, had no bias in checklist-based scoring, indicating that rater 1 had less internal consistency in overall scoring. Perhaps because overall scoring always preceded checklist-based scoring, and rater 1 may have been more familiar with the content of the subject's writing when scoring was done for the second time and had a better understanding of the scoring process. The bias of rater 2 indicated that she lacked a stable standard in the marking process, which shows the importance of rater training, which should make the scorers clear about the scoring descriptors and maintain internal consistency.

The number of biases in dimensions is significantly smaller for overall ratings than for using checklist-based scale, which is caused by the small number of scoring dimensions on the one hand and the granularity of descriptors on the other. For fine-grained scales, it is important to clarify the dimensions of the descriptors in order to achieve both intra-rater and inter-rater consistency. Both checklist-based and overall scores showed large bias in content. This suggests that a clearer delineation of the descriptors of content was not achieved before scoring, and that individual raters had different criteria for the descriptors, which resulted in bias in content dimension.

4. CONCLUSION

MFRM provides a wealth of information to enable testers to gain a penetrating and accurate insight into the entire marking process and the dimensions of testing. With the help of the MFRM, the following findings in terms of rater impact and scoring criteria have emerged from this study. First, the severity levels of raters varied significantly, with rater 5 performing the most severely in both global and descriptor-based scoring. Second, rater 9 and 2 showed overfitting and underfitting in both global and descriptor-based scoring when they examined different writing assignments, indicating that rater 9 had a substantial halo effect while rater 2 had poor internal consistency. Also, when comparing the two distinct rating methods used by the same rater, the absolute values of Z for rater 7 and rater 5 were greater than those for rater 2 in descriptor-based rating and were less than those for rater 2 in the global scoring, indicating that these two raters were more severe when using the descriptor-based checklist and more lenient when using the global scoring. For both rater 9 and 2, they had absolute values of Z greater than 2 in both scoring methods, indicating that they applied the scales excessively and severely for both scoring systems. There is also a substantial variation in the severity levels of the raters but no significant differences in the difficulty of the global scoring dimensions.

Overall, the descriptor-based scoring results fit the model satisfactorily, demonstrating the high validity of the criterion and the reliability of the scoring results. However, due to the limitations of the experimental conditions, the number and representativeness of the research participants in this investigation were insufficient, the majority of whom were new and inexperienced in writing scoring so that they may exert a negative impact on the quality of scoring. On the one hand, the university should provide targeted and specialized training for the scorers, and on the other hand, audio thinking can be employed to analyze the scores in depth in order to reveal the deeper reasons and further improve the scoring criteria.

REFERENCES

[1] Bachman, L. F. Fundamental Considerations in Language Testing [M]. Oxford: Oxford University Press, 1990.

[2] Sawaki, Y. Stricker, L. J. & A. H. Oranje. Factor structure of the TOEFL internet-based test[J]. Language Testing,2009(1): 5-30.

[3] In’nami, Y. & R. Koizumi. Factor structure of the revise TOEIC test: A multi sample analysis[J]. Language Testing,2011(1): 131-152.

[4] Qing-Hua, L. A comparative study of itemized and holistic scoring criteria for TEM4 writing [J]. Foreign Language Testing and Teaching,2014(3): 11-20.

[5] Myford, C. M. & E. W. Wolfe. Understanding Rasch measurement: Detecting and measuring rater effects using many-facet Rasch measurement: Part I[J]. Journal of Applied Measurement,2003(4): 386- 422.

[6] Linacre, J. M. Many-Facet Rasch Measurement[M]. Chicago: MESA Press, 1994.

[7] Bonk, W. J. & G. J. Ockey. A many-facet Rasch analysis of the second language group oral discussion task[J]. Language Testing, 2003 (1): 89-110.

[8] Myford, C. M. & E. W. Wolfe. Monitoring Sources of Variability within the Test of Spoken English Assessment System[R] (TOEFL Research Report NO. 65). Princeton, NJ: Educational Testing Service, 2000.

[9] Ying, X. A study on the influence of personalized feedback from raters on the decision making of CET4 essay raters[J]. Foreign Language Testing and Teaching,2015(1): 1-11.

[10] Park, T. An investigation of an ESL placement test of writing using many-facet Rasch measurement [R/OL]. Teahers College, Working paper in TESOL & Applied Linguistics, Columbia University Academic Commons, 2004(1): 1-23.

[11] Bond, T. G. & C. M. Fox. Applying the Rasch Model: Fundamental Measurement in the Human Sciences[M]. New Jersey: Lawrence Erlbaum Associates, 2015.

[12] Dian, S., Jia-Min, X. A review of research on the quality of assessment tools based on the Rasch model [J]. China Examinations,2020(2):65-71.