Practice-Oriented Paper

Investigating the Rubric for a Writing Class Using Multi-Facets Rasch Analysis

James Horvat

Kwansei Gakuin University

Analytical rubrics are commonly used to assess university students' writing skills; however, rubric validity (content of the rubric and rater interpretation) is a common concern. A rubric developed for use in a coordinated writing program at a large university in western Japan was investigated in spring 2022 for validity. Five raters evaluated the first and second drafts of student essays ($N = 14$). The raters' scores were analyzed using many-facets Rasch software. Subsequently, raters were asked to explain their scoring in individual interviews. The results showed that raters' scores were not consistent with the overall model. Of the five raters, only two had acceptable infit mean squares (1.05 and 0.79 vs. 0.55, 1.37, and 0.35). Additionally, the categories of Topic Sentence and Language/Format showed rater bias in the case of three out of the five raters (18 and 26 instances, respectively). During the interviews, the raters expressed diverging ideas of what constituted an effective topic sentence and to what degree lexical/grammatical errors affected readability. A summary of these results was provided to the program coordinators. After some discussion, a new rubric has been trialed for this coordinated writing program for the 2023 year.

This paper reports the findings of an exploratory study relating to rater behavior and the validity of a rubric for a university writing class. With this study, I aimed to explore raters' perceptions of using the rubric to assess student writing. I used a mixed methods approach that included insights from a many-facets Rasch analysis and rater interviews.

Fulcher and Davidson (2007) defined validity as reaching reasonable conclusions based on the inherent traits of the person being evaluated. Moreover, they suggested validity is circumstantial, as traits are assessed in a certain circumstance, with the assessed person's behavior appropriate for that

circumstance.

Messick (1995) outlined six areas that must be considered when evaluating the validity of an assessment instrument. These areas were content validity, substantive validity, structural validity, generalizability, external validity, and consequence validity. I will discuss content validity and substantive validity the most in this paper. Content validity concerns whether the content of the instrument matches the associated skills and knowledge to be evaluated. The basic construct, including its characteristics and limits, must be reflected in the instrument. Substantive validity means the response and behavior of the test taker is related to the content, or more broadly, the construct to be assessed.

In a pioneering study, Engelhard (1992) examined the ratings of 1,000 randomly chosen eighth-grade essay writing tests by 82 raters. These were high stakes tests required by a southeastern U.S. state. The raters used an analytical rubric with five scoring categories: content/organization, style, sentence formation, usage, and mechanics. Engelhard examined rater severity, relative difficulty of the writing tasks, and overall fit of raters to the model using many-facets Rasch analysis. Despite measures to ensure rater reliability, a large difference emerged in rater severity, with seven misfitting raters showing outfit mean scores of <0.6 or >1.5. Additionally, rater severity ranged from -1.74 to 1.78 logits, and the overall severity differed significantly ($p < .01$). Nevertheless, the five domains of the analytical rubric fit the model with infit mean scores of 0.9 to 1.1. Engelhard's research (1992) spoke to the challenges of rater reliability.

Wolfe and Smith (2007) gave detailed recommendations for assuring studies using Rasch analysis have reliability. These included examining the construct closely and ensuring data are unidimensional, confirming rater and test-taker invariance, and using item-measure correlation and other Rasch analysis output. All these elements showed a connection between the trait or construct to be measured and the instrument to be used. My study examines how clear the construct is in the minds of the raters as they read and apply the rubric.

Lukácsi (2021) examined the validity of an instrument to determine if writing samples were at the CEFR B2 (upper intermediate level, Council of Europe, 2001). A total of 36 questions based on information from the CEFR

B2 level were used to compile a decision-based scale and evaluate four samples. Following the initial review, the scale was then used to evaluate 30 student papers and then 150. The results of a many-facets Rasch analysis showed very good infit mean squares (the majority being around 1.0). This showed a process by which the validity of an instrument could be demonstrated and supported by having a clear construct, a relatively limited area for the rubric, and discussion among the raters.

The present study is a partial replication of Engelhard (1992). I used those procedures and Rasch measures to investigate the rubric used in of the EC3 Writing course at a large private university in western Japan. There was no *a priori* hypothesis.

Research Questions

1. How does overall severity differ between the five raters using the current rubric?
2. How does each rater fit the overall model? Is there evidence of rater bias?
3. How does each student fit the overall model?
4. Are all areas of the rubric being used? Or are most of the scores centered on certain numbers?
5. How do raters perceive the clarity and ease of use of the rubric?

Methods

Participants

These included an intact class taught by me of six male and nine female participants, all first-year university students (aged 18–20 years) majoring in applied social science at a large private university in western Japan. As part of a regular assignment, they wrote two drafts of an essay. Across the 25 classes of the EC3 Writing course, students were placed in groups with similar TOEFL placement test scores. My estimate of their CEFR writing proficiency was B1. EC3 Writing is a coordinated writing course that all students in the faculty needed to take, part of a required two-year, four-skills English program.

The English Language Education Committee of the faculty approved the

procedures for this study. Following the committee's guidelines, all participants signed an informed consent form that detailed the purpose of the study, research methods, and that participation was voluntary. The current course coordinator and program director were informed of the research.

Instrument

All of the instructors teaching the coordinated writing classes use a 10-point analytical rubric scale split into five categories (Topic Sentence, Support 1, Support 2, Conclusion, Language/Format) to assess student performance (Appendix). The rubric ratings as they are written tend to locate and place an average performance as a 7 out of 10. At the time of this study, the scale had been in use for several years, but the developer is unknown (L. Dickerson, personal communication, September 2023).

Procedures

I collected two drafts of short, persuasive student essays from 14 students between May and June 2022. I recruited raters (Table 1) mostly from other teachers in this writing course. Students were provided written feedback in English on the first draft detailing errors of language, organization, and content, then asked to revise. The essays, rater scoring, and subsequent Rasch analysis comprised one data set, while rater interviews, including think-aloud protocols, comprised a second data set. I also rated all of the papers. Two raters rated papers 1 to 7, while two other raters rated papers 8 to 14 (Table 2). Each rater scored both Draft 1 and Draft 2. Raters received the rubric copy, paper format and organization information, and a copy of the task. All essays were anonymized before raters received copies. The scores were analyzed by FACETS software (Linacre, 2022), and Figures 1 to 6 in this paper were output by the software. All categories of the rubric were anchored at 0.

After the Rasch output was analyzed, interviews with the raters were conducted. Think-aloud protocols were used. In these protocols, I asked raters to explain why they assigned a particular score to a certain student. I especially asked about students who did not fit the Rasch model or criteria areas which had high rater bias, especially Language / Format.

Table 1

Rater Biographical Details

Rater	Years in Japan	Years in school system in Japan	Years teaching writing in Japan	Highest degree	Pursuing Ph.D?	Nationality
1	25	15	15	M.A.	In progress	American
2	16	13	8	M.A.	In progress	Canadian
3	8	8	0	M.A.	In progress	American
4	10	10	7	M.A.	In progress	British
5	20	17	3	M.A.	No	Australian

Table 2

Rater Biographical Details

Rater	Student Essays Scored
1	1-14
2	1-7
3	1-7
4	8-15
5	8-15

Results

Wright Map and Statistics

Figure 1 shows the Wright Map. There is a range of student ability with 4 as the highest and 2 as the lowest. There are also differences in rater severity, with Rater 3 being most severe and Rater 1 being the most lenient. Draft 1 and 2 are of equal difficulty. For rubric criteria, Conclusion and Topic Sentence are the most difficult, while Support 1 and Support 2 are the least difficult.

I observed only minor difference between the scores for the two drafts (Figure 2). Infit mean-square results were .94 and .90 for Drafts 1 and 2, respectively.
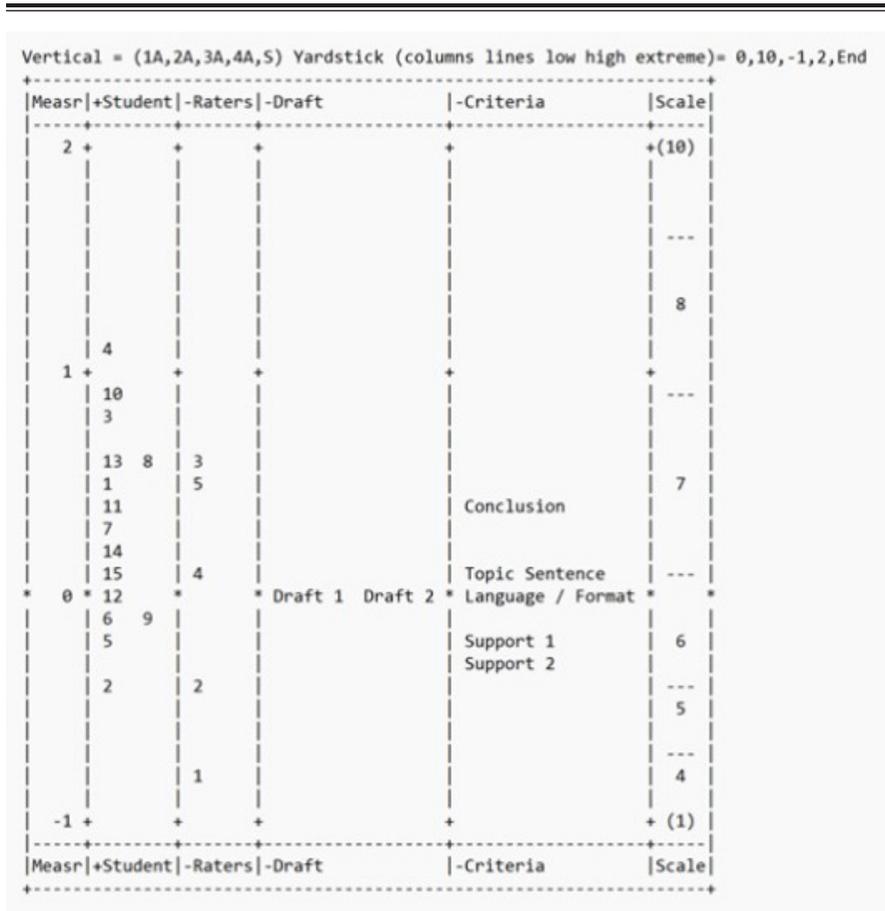
```
Vertical = (1A,2A,3A,4A,S) Yardstick (columns lines low high extreme)= 0,10,-1,2,End
+-------------------------------------------------------------------------+
|Measr|+Student|-Raters|-Draft            |-Criteria           |Scale|
|-----+--------+-------+------------------+--------------------+-----|
|  2 +         +       +                  +                    +(10) |
|    |         |       |                  |                    |     |
|    |         |       |                  |                    |     |
|    |         |       |                  |                    | --- |
|    |         |       |                  |                    |     |
|    |         |       |                  |                    |  8  |
|    | 4       |       |                  |                    |     |
|  1 +         +       +                  +                    +     |
|    | 10      |       |                  |                    | --- |
|    | 3       |       |                  |                    |     |
|    | 13  8   | 3     |                  |                    |     |
|    | 1       | 5     |                  |                    |  7  |
|    | 11      |       |                  | Conclusion         |     |
|    | 7       |       |                  |                    |     |
|    | 14      |       |                  |                    |     |
|    | 15      | 4     |                  | Topic Sentence     | --- |
|  * 0 * 12    *       * Draft 1  Draft 2 * Language / Format  *     *
|    | 6   9   |       |                  |                    |     |
|    | 5       |       |                  | Support 1          |  6  |
|    |         |       |                  | Support 2          |     |
|    | 2       | 2     |                  |                    | --- |
|    |         |       |                  |                    |  5  |
|    |         |       |                  |                    | --- |
|    |         | 1     |                  |                    |  4  |
|    |         |       |                  |                    |     |
| -1 +         +       +                  +                    + (1) |
|-----+--------+-------+------------------+--------------------+-----|
|Measr|+Student|-Raters|-Draft            |-Criteria           |Scale|
+-------------------------------------------------------------------------+
```

Figure 1. Wright Map.

```
+-----------------------------------------------------------------------------------------------+
| Total  Total   Obsvd Fair(M)|  -   Model | Infit      Outfit     |Estim.|      | Correlation  |
| Score  Count Average Average|Measure S.E. | MnSq ZStd  MnSq ZStd |Discrm| Displ.| PtMea PtExp | N Draft
|-----------------------------------------------------------------------------------------------|
| 1477    225    6.56   6.79 A|  .00   .05 | .94 -.5    .99  .0 | 1.06 |  .11 |  .62   .67 | 1 Draft 1
| 1564    225    6.95   6.79 A|  .00   .05 | .90 -.9    .92 -.7 |  .95 | -.11 |  .72   .67 | 2 Draft 2
|-----------------------------------------------------------------------------------------------|
| 1520.5  225.0  6.76   6.79 |   .00   .05 | .92 -.8    .95 -.4 |      |      |  .67 | Mean (Count: 2)
|   43.5    .0    .19    .00 |   .00   .00 | .02  .2    .03  .3 |      |      |  .05 | S.D. (Population)
|   61.5    .0    .27    .00 |   .00   .00 | .03  .3    .05  .4 |      |      |  .07 | S.D. (Sample)
+-----------------------------------------------------------------------------------------------+
Model, Populn: RMSE .05  Adj (True) S.D. .00  Separation .00  Strata .33  Reliability .00
Model, Sample: RMSE .05  Adj (True) S.D. .00  Separation .00  Strata .33  Reliability .00
Model, Fixed (all same) chi-square:  .0  d.f.: 1  significance (probability): 1.00
```

Figure 2. Statistics are listed such as infit mean square and outfit mean square for Draft 1 and Draft 2.

Thus, infit statistics were good, as a score of .70 to 1.30 shows a relatively good fit to the model (Linacre, 2023; Wright & Linacre, 1994). Average scores for Draft 2 were higher (6.95) than Draft 1 (6.56), suggesting some improvement.

Concerning the raters (Figure 3), two of the five raters showed a good fit to the model. Rater 1 showed an infit mean-square of 1.05, and Rater 5 showed .79. However, Raters 2, 3, and 4 produced infit mean-square results of .55, 1.57, and .35, respectively. Unfortunately, the results indicate these three raters did not fit the model.

I also noted differences in rater severity. Raters 3 and 5 were the strictest, giving an average score of 5.41 and 5.84 across the five 10-point categories, respectively. Raters 2 and 4 were in the middle with averages of 7.16 and 6.51, respectively, across the five categories. Rater 1 was the most lenient with an average of 7.82.

The student report was omitted to save space. Of the 14 students, the observed average (combining all scores) data was (MIN = 5.53, MAX = 8.07, *M* = 6.76). The fair average data was (MIN = 5.36, MAX = 7.74, *M* = 6.70). Infit mean square data was (MIN = .32, MAX = 2.61, *M* = .96). 7 of the 14 students fit the model, with InfitMS in the range from 0.70 to 1.30.

The results demonstrated each category had some bias (Table 3), which was determined by tallying and summarizing the data from the Rater Bias Report, as output by the FACETS software (Linacre, 2022). Topic Sentence, Support 2, and Conclusion had similar numbers of bias reports (M = 25). Language/format

```
+---------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair(M)|   -    Model | Infit      Outfit     |Estim.| Correlation |          |
| Score   Count  Average Average|Measure S.E.  | MnSq ZStd  MnSq ZStd  |Discrm| PtMea PtExp | N Raters |
|---------------------------------------------------------------------------------------------------|
|  379     70    5.41    5.78  |  .56    .08  | 1.57  2.9  1.98  4.2  |  .82 |  .79   .65  | 3 3      |
|  467     80    5.84    5.94  |  .49    .07  |  .79 -1.2   .83  -.9  |  .78 |  .67   .51  | 5 5      |
|  521     80    6.51    6.58  |  .15    .09  |  .35 -4.5   .33 -4.7  | 1.24 |  .64   .45  | 4 4      |
|  501     70    7.16    7.23  | -.36    .10  |  .55 -2.6   .57 -2.4  |  .94 |  .54   .53  | 2 2      |
| 1173    150    7.82    7.78  | -.83    .07  | 1.05   .4  1.06   .4  | 1.09 |  .12   .48  | 1 1      |
|---------------------------------------------------------------------------------------------------|
|  608.2   90.0  6.55    6.67  |  .00    .08  |  .86 -1.0   .95  -.7  |      |  .55        | Mean (Count: 5)   |
|  286.6   30.3   .87     .76  |  .53    .01  |  .42  2.6   .57  3.0  |      |  .23        | S.D. (Population)  |
|  320.4   33.9   .97     .85  |  .59    .01  |  .47  2.9   .64  3.4  |      |  .26        | S.D. (Sample)      |
+---------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .08  Adj (True) S.D. .52  Separation 6.33  Strata 8.78  Reliability .98
Model, Sample: RMSE .08  Adj (True) S.D. .58  Separation 7.10  Strata 9.80  Reliability .98
Model, Fixed (all same) chi-square: 245.1  d.f.: 4  significance (probability): .00
Model, Random (normal) chi-square: 3.9  d.f.: 3  significance (probability): .27
---------------------------------------------------------------------------------------------------
```

Figure 3. Rater information for Raters 1 to 5, including Obsvd Average (average score of the five categories) and infit mean square.

Table 3

Rater Bias Totals

Rater	Topic sentence	Support 1	Support 2	Conclusion	Language/Format
1	3	6	14	7	12
2	3	0	3	6	9
3	5	3	5	2	5
4	6	2	3	6	4
5	7	4	1	5	4
Total	24	15	26	25	34

had the most bias at 34. Rater 1's totals were nearly double those of the other raters in three aspects of the rubric, since Rater 1 rated twice as many essays.

Turning to the criteria items on the rubric, Topic Sentence, Language/Format, and Support 2 showed a good fit to the model (Figure 4) with infit mean results of 1.05, 1.19, and 1.03, respectively. Support 1 had a somewhat poor fit at .62 with an infit Z score of -2.56. The category Conclusion produced an infit mean result of 0.72 and an infit Z score of -1.89. These numbers appear to contrast with rater bias information, in which Topic Sentence and Language/Format appeared prominently in the rater bias figures. Overall, four of the rating

```
+------------------------------------------------------------------------------------+
| Total   Total  Obsvd  Fair(M)|   -    Model | Infit     Outfit    |Estim.| Correlation |        |
| Score   Count Average Average|Measure  S.E. | MnSq ZStd MnSq ZStd |Discrm| PtMea PtExp | N Criteria |
|-------------------------------------------+----------+-------------+--------------------|
|  550      90    6.11   6.22 |   .35    .07 |  .72 -1.8  .78 -1.3 | 1.07 |  .80   .68 | 4 Conclusion      |
|  595      90    6.61   6.66 |   .09    .08 | 1.05   .3 1.03   .2 |  .87 |  .58   .66 | 1 Topic Sentence  |
|  608      90    6.76   6.78 |   .01    .08 | 1.19  1.1 1.23  1.3 |  .66 |  .59   .65 | 5 Language / Format |
|  634      90    7.04   7.00 |  -.16    .08 |  .62 -2.5  .70 -1.9 | 1.29 |  .62   .64 | 2 Support 1       |
|  654      90    7.27   7.17 |  -.30    .09 | 1.03   .2 1.03   .2 | 1.12 |  .66   .64 | 3 Support 2       |
|-------------------------------------------+----------+-------------+--------------------|
|  608.2    90.0  6.76   6.76 |   .00    .08 |  .92  -.6  .95  -.3 |      |        .65 | Mean (Count: 5)   |
|   35.6      .0   .40    .32 |   .22    .00 |  .21  1.4  .19  1.2 |      |        .08 | S.D. (Population) |
|   39.8      .0   .44    .36 |   .25    .00 |  .24  1.6  .21  1.3 |      |        .09 | S.D. (Sample)     |
+------------------------------------------------------------------------------------+
Model, Populn: RMSE .08  Adj (True) S.D. .21  Separation 2.61  Strata 3.81  Reliability .87
Model, Sample: RMSE .08  Adj (True) S.D. .24  Separation 2.96  Strata 4.28  Reliability .90
Model, Fixed (all same) chi-square:  40.2  d.f.: 4  significance (probability): .00
Model, Random (normal) chi-square:   3.6  d.f.: 3  significance (probability): .30
------------------------------------------------------------------------------------
```

Figure 4. The five criteria items of the rubric are shown with fit statistics and the average score for each item (Obsvd Average).

categories produced values within the .70 to 1.30 range for infit mean square, which indicates the rubric categories mostly fit the model.

When examining whether all 10 ranking levels of the rubric (Figure 5) were used for each draft, I found that rankings 1, 2, and 3 were used the least overall (<4% of the time). Similarly, rankings 9 and 10 were also used infrequently (<11% of the time). However, rankings 6, 7, and 8 were used 71% of the time. It appears that more areas of the ranking scale could have been used. Gaps in the rubric, where there is no description, probably contributed to rater uncertainty as to the meaning of those ranking levels and subsequent unwillingness to use those (please see Appendix).

The Rasch-Andrich Thresholds (Figure 5) are neither consistently in order from highest to lowest, nor consistently in order from lowest to highest. These sequential sequences would indicate that all areas of the scale are being used. Instead, the numbers alternate between positive and negative figures. Therefore, it appears that not all areas of the rating scale were used.

Rater Interviews

Based on the rater interview data, it was clear that they were confused about the rubric items. On the surface, the analytical rubric with 5 categories appears to divide the task of evaluation among the five criteria of Topic Sentence, Support 1, Support 2, Conclusion, and Language/Format. An impression is given of 5 criteria areas, each with distinct and well-defined ranking bands, from 1 to 10. However, further investigation showed that the raters did not appear to have a

```
Model = ?B,?B,?B,?B,R10
+----------------------------------------------------------------------------------------------------+
|        DATA                    |    QUALITY CONTROL |RASCH-ANDRICH|  EXPECTATION  |  MOST  |  RASCH- | Cat|
|   Category Counts      Cum.|  Avge  Exp. OUTFIT| Thresholds  |   Measure at  |PROBABLE| THURSTONE|PEAK|
|Score Total    Used   %   % |  Meas  Meas  MnSq |Measure S.E.|Category  -0.5 |  from  |Thresholds|Prob|
|--------------------------------+-------------------+-------------+---------------+--------+----------+----|
|  1     10      10    2%   2% | -.99  -.80   .5 |             |( -2.10)       |  low   |   low    |100%|
|  2      2       2    0%   3% | -.68  -.67   .7 |  .88    .35| -1.49  -1.80|        |  -1.30   |  6%|
|  3      7       7    2%   4% | -.56  -.52   .6 | -1.85   .33| -1.13  -1.29|        |  -1.23   | 10%|
|  4     42      42    9%  14% | -.49  -.33   .5 | -2.22   .27|  -.84   -.98| -1.06  |  -1.10   | 35%|
|  5     25      25    6%  19% | -.11  -.12   .8 |  .29    .17|  -.54   -.70|        |   -.63   | 13%|
|  6     61      61   14%  33% |  .07   .15   .7 | -.88    .15|  -.16   -.37|        |   -.45   | 22%|
|  7    157     157   35%  68% |  .55   .46  1.1 | -.64    .12|   .46    .10|  -.41  |   -.11   | 47%|
|  8    100     100   22%  90% | 1.00   .81   .7 | 1.09    .12|  1.29    .89|  1.09  |    .98   | 39%|
|  9     12      12    3%  92% |  .84* 1.14  1.5 | 3.10    .18|  1.99   1.64|        |   1.79   |  9%|
| 10     34      34    8% 100% |  .93  1.40  1.7 |  .24    .20|( 2.89)  2.46|  1.67  |   1.95   |100%|
+---------------------------------------------------------------(Mean)--------(Modal)--(Median)------+
```

Figure 5. Category statistics for the individual numbers of the rating scale, including percentage used.
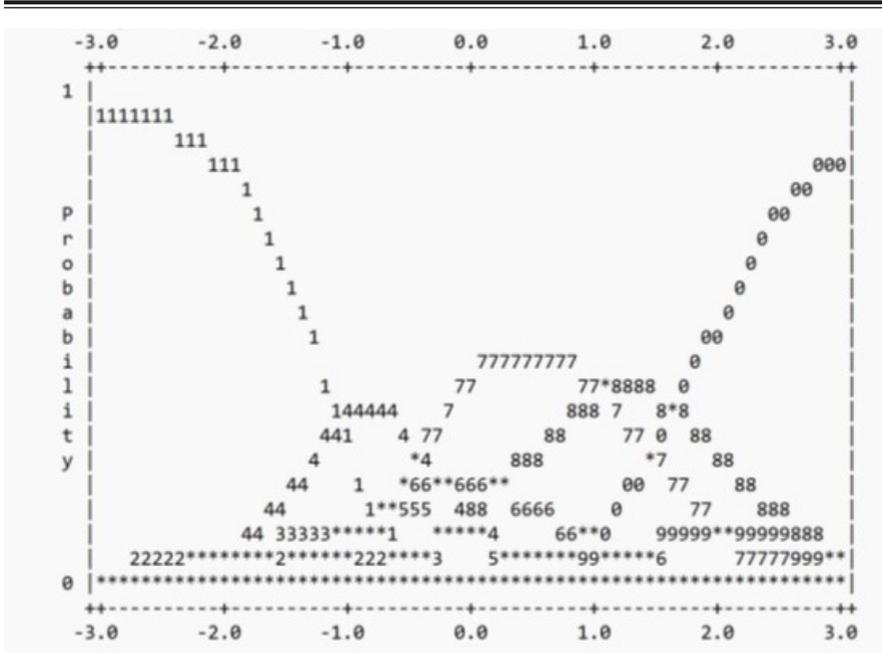
Figure 4. Probability curves showing the use of the range of the rating scale. Full-bodied curves, similar to sine waves, indicate that all areas of the rating scale are being used. Curves that appear to be flattened, or less than full-bodied, indicate that some areas of the rating scale are not being used.

clear idea of the meaning of some of the criteria items in the rubric. For example, two raters stated they believed the topic sentence needed to be original or "novel" to gain a high score, whereas another rater stated they understood that the topic sentence only needed to summarize what would come later in the paragraph. Then, it appears that Raters 4 and 5 and Rater 1 had a different understanding. Topic Sentence Ranking 7 indicates it needs to introduce "the main subject with a controlling idea that adequately guides the rest of the paragraph", and for Ranking 10, it needs to introduce the "main subject and has a clear controlling idea that guides the rest of the paragraph exceptionally well." Originality is not mentioned.

Similarly, the raters did not appear to agree on how to score the Language/Format category. Agreement was lacking on which types of grammar mistakes were, nonetheless, comprehensible to the person rating the paper. Additionally,

some raters appeared to bring into Language/Format items not explicitly stated in the rubric.

For example, when discussing accuracy, Rater 2 stated, "I am compelled to find every single spelling and grammar mistake there is and to make sure the students are aware of those". When I asked, "Do you feel you might be a little harsher, a little more conscientious, or thorough than other teachers?" Rater 2 answered, "All of the above."

Rater 4 interpreted the rubric differently than Raters 2 and 3 in terms of the degree to which grammar issues affect readability, "...for the most part, it's grammatically pretty good. There's no major spelling errors or word choices that strain the reader." The notion that some errors do not strain the reader very much appears not to be shared by the raters, and subsequently they did not have the same construct regarding accuracy of language as they rated. The rubric for Language/Format states, "There are grammar errors that considerably affect readability" for ranking 4, "There may be some grammar errors that slightly affect readability but is understandable without too much effort" for ranking 7, and "There are few, if any, grammar errors" for ranking 10. This may help explain why the Language/Format area (Table 3) for rater bias was the highest number and why some of the raters did not fit the model, as shown by their Infit Mean-Square statistics (Figure 3).

Discussion

We can see differences in the raters' interpretations of the rubric. This is due to a lack of agreement on the constructs the rubric aims to assess. Returning to Messick's (1995) concepts of validity, content validity is the connection between the knowledge and skills being assessed and the instrument's representation of these. Some of the raters believed that originality was necessary for a good topic sentence, but some did not believe this. Some of the raters believed that even minor grammatical errors greatly reduced reader comprehensibility of a sentence, while others believed that the grammar errors had to be prominent to reduce comprehensibility. It appears that the lack of clarity in the raters' understanding of the criteria items stems from the rubric. The infit statistics of the raters indicate

that the scores of the same student over several ratings by different raters were not very consistent, which relates to structural validity (Messick, 1995). Indeed, as discussed previously, Wolfe and Smith (2007) recommended that the construct be clear to be evaluated. If the construct is clear, and student ability level remains relatively constant, invariance will be present and consistent ratings should be given.

Based on the findings of this study, the current rubric requires revision. First, examples for each level of performance would help to add clarity. For example, topic sentences demonstrating poor, weak, satisfactory, and excellent performance should be included. As indicated in the interview data, raters expressed confusion regarding how to rate the novelty or originality of topic sentences. If raters are provided with examples of expected performance and opportunities to practice scoring, they will have a better understanding of the rating categories, rankings, and the constructs represented by those categories. These changes would improve content and substantive validity, as defined by Messick (1995).

Concerning the issues raised with the Language/Format category, Format should be a separate category. The Format category would include correct font type and size, double-spacing of the text, paragraph indentation, etc. Other aspects of the category would include academic standards, such as avoiding contractions. Under the category of Language, concrete examples of sentences containing errors, but at different levels of readability, should be provided. This would aid in having a relatively clear construct and a relatively limited area covered by the rubric for each category that Lukácsi (2021) recommended.

I must also acknowledge the limitations of this study. First, the sample size of students was small ($N = 14$). Another limitation is that their TOEFL scores were high, compared with the other students in EC3 Writing. Examining a middle and a lower-level class would have been helpful. Another limitation is that it was not possible to interview the person who created the rubric, as their identity was unknown. Additionally, it is a bit surprising in my study that many raters and students did not fit the model, as indicated by their infit mean scores, but that the individual rubric items, such as Topic Sentence, did fit the model.

Engelhard (1992) had a similar result in his study.

Conclusion

With this study, I reviewed a criteria scale to examine construct validity. Overall, the results indicated that the rating scale had specific weaknesses in terms of effectiveness. This is shown in infit mean square statistics of the raters and in the rater bias reports.

References

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9781410614575

Council of Europe. (2001). C*ommon European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Engelhard G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171–191. https://doi.org/10.1207/s15324818ame0503_1

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced reference book*. Routledge.

Linacre, J. M. (2023). Facets computer program for many-facet Rasch measurement (Version 3.85.1) [Computer software]. Winsteps. https://www.winsteps.com/facets.htm

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing*, *38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Wolfe, E., & Smith, E. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, *8*(2), 204–234. PMID: 17215568.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values.

Rasch Measurement Transactions, 8(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Author Bio

James Horvat works as an instructor at three large private universities in western Japan. He holds a Master's degree in Applied Linguistics from the University of Southern Queensland, Australia. He is a PhD student in Applied Linguistics at Temple University Japan. *jhorvat111@kwansei.ac.jp*

Appendix

	Poor		Weak				Satisfactory			Excellent
	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Topic Sentence (15%)	Lacks clear main topic or is missing the contributing idea		Introduces the main subject with a controlling idea that is somewhat disconnected to the rest of the paragraph.				Introduces the main subject with a controlling idea that adequately guides the rest of the paragraph.			Introduces the main subject and has a clear controlling idea that guides the rest of the paragraph.
Support 1 (25%)	Not given or is unrelated to the topic. Details are too few, not written, or do not help the reader understand the support's connection to the topic sentence.		Somewhat related to the topic. Details are not clearly connected or do not help the reader understand the support's connection to the topic sentence.				Mostly connected to the topic. Details give the reader information that somewhat help in understanding the support's connection to the topic sentence.			Logical and clearly relevant to the topic. Details give more than enough information for the reader to clearly understand the support's connection to the topic sent.
Support 2 (25%)	Not given or is unrelated to the topic. Details are too few, not written, or do not help the reader understand the support's connection to the topic sentence.		Somewhat related to the topic. Details are not clearly connected or do not help the reader understand the support's connection to the topic sentence.				Mostly connected to the topic. Details give the reader information that somewhat help in understanding the support's connection to the topic sentence.			Logical and clearly relevant to the topic. Details give more than enough information for the reader to clearly understand the support's connection to the topic sent.

	Poor	Weak			Satisfactory			Excellent
	1.0 2.0	3.0 4.0	5.0	6.0	7.0	8.0	9.0	10.0
Conclusion (15%)	Missing from the paragraph.	Present, but does not offer a clear connection to the topic.			Adequately refers to the topic, or it offers an opinion, prediction, or suggestion that are somewhat connected to the topic.			Makes it clear that the paragraph is coming to a close. It restates the topic, or it offers a relevant opinion, prediction, or suggestion. This conclusion exceeds expectations
Language / Format (25%)	Grammar errors greatly affect readability. The paragraph lacks a logical sequence, or the writing is off-topic for the assignment. Major formatting or stylistic errors.	Grammar errors that considerably affect readability. The paragraph lacks a logical sequence and is somewhat relevant to the assignment. Multiple Academic Writing and some formatting errors may be present.			May be some grammar errors that slightly affect readability, but it is understandable without too much effort. Paragraph is mostly logically sequenced and relevant to the assignment. Few or no Academic Writing or formatting errors.			Few, if any, grammar errors. The paragraph is written in perfect academic style. It is logically sequenced and all parts are relevant to the assignment, Perfect formatting.