Attractive : Innovative Education Journal
Vol. 6 No. 1, March 2024

Item Fit Analysis for Evaluating Academic Writing Performance With Rasch Measurement

Leny Hikmah Rentiana[1], Ulfa Rahma Dhini[1], Lilik Yuliawati[1], Eka Tri Wulandari[1]
Universitas Bina Sarana Informatika, Indonesia

Corresponding Author: leny.lyh@bsi.ac.id*

ARTICLE INFO
Article history:
Received
Apri; 02, 2024
Revised
April 26, 2024
Accepted
May 09, 2024

ABSTRACT

This study aims to evaluate item fit through the application of the Rasch model. The research involved 40 English as a foreign language student who took part in an essay writing class as part of a TOEFL iBT course, with writing samples gathered in partnership with the language center offering the TOEFL iBT classes. The research methodology utilizes the Rasch model with Winsteps software for quantitative analysis, employing three key statistical outputs: a "statistical summary output" to provide overall data and figures, item statistics to assess item validity, and participant statistics to evaluate participant validity. Based on the interpretation of misfit order Rasch measurement for item fit, four items ("Content," "Structure," "Mechanic," and "Diction") demonstrated excellent alignment with the Rasch model. All items exhibited Outfit Mean Square (MNSQ) values below the expected 1.0 yet within the acceptable range (0.5 - 1.5), indicating better-than-expected fit, with "Diction" showing the highest conformity. The Outfit Z-Standard (ZSTD) values for all items were near zero, signifying no significant deviations from the model. High Point Measure Correlations (PT Measure Corr) of 0.91 for "Content" and "Diction" and 0.88 for "Structure" and "Mechanic" suggest strong consistency with the overall measured ability, affirming the significant and valid contributions of these items to the scale's measurement objectives. Thus, all four items are deemed excellent within the applied measurement scale, particularly "Diction," which exhibits the highest model fit.

Keywords: *Item Fit, Academic Writing, Rasch Measurement*

Journal Homepage https://www.attractivejournal.com/index.php/aj/

Published by CV. Creative Tugu Pena

INTRODUCTION

The evaluation of academic writing proficiency is a cornerstone of educational assessment, playing a pivotal role in gauging students' mastery of course material, critical thinking skills, and ability to communicate effectively. However, traditional methods of assessing academic writing often rely on subjective judgments, such as instructor grading, which can be inconsistent and prone to bias (Meisel et al, 2017). In response to these challenges, the Rasch Measurement offers a promising framework for objective and reliable measurement by analyzing the relationship between item difficulty and test-taker ability (Mccreary et al, 2013). This thesis explores the application of the Rasch model to the assessment of academic writing performance, focusing specifically on item fit analysis as a means of enhancing the validity and precision of writing assessments.

Item fit analysis, a fundamental aspect of Rasch measurement, examines the degree to which individual test items conform to the expectations of the measurement model (Huang, 2023; Stemler and Naples, 2021). In the context of academic writing assessment, item fit analysis involves scrutinizing the appropriateness of writing prompts or tasks in relation to the construct being measured and the abilities of the test-takers(Li, Pan, and Wang 2021). By identifying and addressing discrepancies between expected and observed responses, item fit analysis aims to ensure that assessment instruments accurately reflect students' writing proficiency (Reise, 1990; Rost, J., & von Davier, 1994; Uto, 2022).

One of the key challenges in assessing academic writing using the Rasch model lies in the design and selection of appropriate writing tasks (Erguvan and Aksu Dunya, 2020). Writing prompts must be carefully crafted to align with the intended learning outcomes and to elicit responses that accurately reflect students' writing abilities. Additionally, considerations such as task difficulty, language complexity, and cultural relevance must be taken into account to ensure the validity and fairness of assessments across diverse student populations and contexts.

Moreover, the Rasch model offers unique advantages in addressing some of the limitations of traditional writing assessment methods (Rahman, 2023; Rost, J., & von Davier 1994). By providing a framework for estimating both item difficulty and test-taker ability on a common metric, the Rasch model enables more precise measurement and comparability of writing performance across different assessments and populations. Furthermore, the objectivity and reliability afforded by Rasch measurement can help mitigate the subjectivity and inconsistency inherent in traditional grading practices .

In recent years, there has been growing interest in the application of Rasch analysis to various domains of educational assessment, including academic writing. However, research specifically focused on item fit analysis within the Rasch framework for writing assessment remains relatively limited. This thesis aims to fill this gap by conducting a comprehensive investigation into the utility and implications of item fit analysis for assessing academic writing performance. By examining the alignment between writing prompts and the Rasch measurement model, this study seeks to provide insights that can inform the development of more valid and reliable writing assessments in educational settings. Through empirical analysis and theoretical exploration, this thesis contributes to advancing our understanding of how the Rasch model can be leveraged to enhance the assessment of academic writing proficiency.

METHOD

This study involved 40 EFL students from the Eloquensi English Language Centre who had completed a TOEFL iBT essay writing course. These students, who were in high school or college and had intermediate or higher levels of English proficiency, were instructed to write five-paragraph essays following a standardized format consisting of an introduction, body, and conclusion. This structure was implemented to ensure fair evaluations by the raters, as the number of paragraphs could potentially impact the ratings.

The Rasch measurement model was used for analysis due to its capability to simultaneously capture the interaction between respondents and items. Unlike raw scores, the Rasch model uses logit values to indicate the probability of an item being chosen by a group of respondents (Sumintono, B. & Widhiarso, 2013; Tan, 2013). The Rasch model was employed to estimate expected raw scores for Likert ratings, which

are ordinal and have unequal intervals between scores. The application of the Rasch model to polytomous data was developed by Andrich (1978), based on two key theorems: individual ability/agreement levels and item difficulty levels in achieving agreement (Misbach & Sumintono, 2014). The outputs used for data analysis included summary statistics for reliability (Figure 1), unidimensionality items (Figure 2), and Fit Order items (Figure 3) for validity.

A holistic rubric by Jacob et al (1981)was used to measure writing skills. This rubric includes six levels: proficient, fluent, expanding, developing, beginning, and emerging. It assesses four key elements of writing ability: content, structure, diction, and mechanics. Content evaluation covers the introduction, ideas in body paragraphs, and logical flow of ideas. Structure assessment examines grammar application and sentence variety (simple, compound, complex, and compound-complex sentences). Diction evaluates vocabulary use and variation to avoid repetition. Lastly, mechanics assesses punctuation, spelling, and capitalization accuracy.

Analyzing item fit in the Rasch model involves several systematic steps to ensure the quality and reliability of the assessment tool. The model provides steps to analysing it.

1. Data Preparation and Entry:Collect Responses: Administer the test to a representative sample and collect the responses.Organize Data: Arrange the data in a format suitable for Rasch analysis software, ensuring accurate and complete data entry for each respondent and item.
2. Model Fitting: Choose Software: Select appropriate Rasch analysis software, here Winsteps, Fit the Model: Input the data into the software and run the initial Rasch analysis to fit the model to the collected data. The software will estimate item difficulty and person ability parameters.
3. Evaluate Item Fit Statistics:Infit and Outfit Mean-Square Statistics: Examine the infit and outfit mean-square values for each item. These statistics indicate the degree to which the data fit the Rasch model. Infit: Sensitive to unexpected responses close to the person's ability level. Outfit: Sensitive to unexpected responses far from the person's ability level.

RESULTS AND DISCUSSION

Three output data results are used to reveal the reliability of items. They are summary of four measured statistics, Items Statistics, and misfit order. Summary of four measured item displays the general outcome of the four measured item which consist infit and outfit plus the MNSQ and ZSTD score. Misfit order record the more specific items which contribute to inform the score number of the four items: content, structure, mechanics, and diction.

```
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .92  SEM = .79
STANDARDIZED (50 ITEM) RELIABILITY = .95

     SUMMARY OF 4 MEASURED ITEM
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     141.8      40.0         .00     .47       .96   -.09    .57    .04 |
|  SEM       2.4        .0         .50     .01       .10    .39    .05    .06 |
| P.SD       4.1        .0         .86     .02       .17    .67    .10    .11 |
| S.SD       4.7        .0         .99     .02       .20    .78    .11    .13 |
| MAX.     146.0      40.0        1.41     .50      1.22    .92    .70    .23 |
| MIN.     135.0      40.0        -.93     .46       .78   -.89    .44   -.03 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .49 TRUE SD     .71  SEPARATION  1.46  ITEM   RELIABILITY  .68 |
|MODEL RMSE    .47 TRUE SD     .72  SEPARATION  1.52  ITEM   RELIABILITY  .70 |
| S.E. OF ITEM MEAN = .50                                                     |
-------------------------------------------------------------------------------
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000
```

Figure 1. Summary of Item Reliability

It is important to convey that Cronbach's alpha is used to measure the reliability of items, representing the overall interaction between persons and items. Additionally, INFIT MNSQ and OUTFIT MNSQ values can be utilized, with values closer to 1.00 indicating better fit. Similarly, the ideal values for INFIT ZSTD and OUTFIT ZSTD are 0.0, meaning the closer these values are to 0.0, the higher the quality (Linacre, 2002; Sumintono, B. & Widhiarso, 2013). In summary, the interpretation of the two sets of parameters can be outlined as follows:

1. Infit MNSQ and Outfit MNSQ: The expected value is around 1. Values between 0.5 and 1.5 are typically considered indicative of adequate fit. Values greater than 1.5 indicate overfit (data too closely matches the model) or underfit (data does not match the model), depending on the context and application.
2. Z-Standardized Fit (Infit ZSTD and Outfit ZSTD): The expected z-standard value is around 0, with values between -2 and +2 generally considered acceptable.

In this context, the Infit Mean Square (Infit MNSQ) measures how well respondents near the item's ability level respond according to the Rasch model. Meanwhile, the Outfit Mean Square (Outfit MNSQ) measures how well all respondents (including those far from the item's ability level) respond according to the Rasch model.

Figure 1 provides information on four items. The first data displayed is the standard error of the item (S.E. OF ITEM MEAN), which is 0.50. The second noted item is the item separation index, which is 1.46. Additionally, above the table, there is information on Cronbach's alpha value, which is 0.92. Finally, the item reliability value is 0.68. The interpretation of the summary statistics from the Rasch measurement for item fit is as follows:

1. Separation (1.46): In Rasch measurement, separation indicates how well the scale can differentiate between items of varying difficulty. A separation value of 1.46 suggests that the scale's ability to distinguish between items is adequate but not

optimal. Generally, a higher separation value (greater than 2) is desired to show that the scale has good discriminative power.

2. Infit MNSQ (0.96) and Infit ZSTD (-0.09): An Infit Mean Square (MNSQ) value of 0.96 indicates that the items have a good fit with the Rasch model. The expected value for Infit MNSQ is around 1.0, with an acceptable range typically between 0.7 and 1.3. An Infit ZSTD (standardized fit statistic) value of -0.09, which is close to 0, suggests no significant deviation from the Rasch model.
3. Outfit MNSQ (0.57) and Outfit ZSTD (0.04): An Outfit Mean Square (MNSQ) value of 0.57, though lower than the expected value of 1.0, is still within the tolerance range of 0.5 to 1.5. This might indicate that some items are more predictive or fit the model better than others. An Outfit ZSTD value of 0.04, which is close to 0, also indicates no significant deviation in outfit fit.
4. Reliability (0.68): A reliability value of 0.68 suggests a moderate level of reliability for the scale. This indicates moderate consistency in measurement results but is not high enough to be considered very reliable. For better measurement, a reliability value above 0.8 is typically desired.
5. Standard Error of Item Mean (0.50): A Standard Error of Item Mean (S.E. of Item Mean) of 0.50 indicates the precision of the estimated mean item difficulty. Smaller SE values indicate more precise estimates. A value of 0.50 suggests moderate uncertainty in the estimation of average item difficulty.

Overall, these results indicate that the items on the scale have a reasonably good fit with the Rasch model, though there are areas for improvement, such as increasing separation and reliability to produce a more discriminative and reliable scale.

In Rasch analysis, "misfit order" refers to data behavior that does not align with the expected Rasch model. The Rasch model aims to measure an individual's ability or an item's difficulty level based on the given responses. Misfit can be interpreted in several ways depending on the type of misfit occurring. Two statistics commonly used to evaluate misfit are Outfit Mean Square (Outfit MNSQ) and Infit Mean Square (Infit MNSQ). Outfit MNSQ: Sensitive to outliers or unusual responses on items that are very easy or very difficult. Infit MNSQ: Sensitive to misfit occurring on items with difficulty levels matching the participants' abilities (items around the participants' ability level). Both statistics typically have an expected value of 1. Values significantly higher or lower than 1 indicate misfit. The general interpretation is as follows:

1. MNSQ values between 0.5 and 1.5 are considered reasonable fit.
2. MNSQ values below 0.5 indicate data that fits the model too well (overfit).
3. MNSQ values above 1.5 indicate data that does not fit the model (misfit).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
1	143	40	-.24	.47	1.22	.92	.70	.23	A .91	.89	80.0	83.4	Content
2	135	40	1.41	.46	1.00	.06	.61	-.03	B .88	.88	85.0	81.4	Structure
4	146	40	-.93	.50	.84	-.45	.53	-.01	b .88	.89	87.5	86.5	Mechanic
3	143	40	-.24	.47	.78	-.89	.44	-.03	a .91	.89	85.0	83.4	Diction
MEAN	141.8	40.0	.00	.47	.96	-.1	.57	.0			84.4	83.7	
P.SD	4.1	.0	.86	.02	.17	.7	.10	.1			2.7	1.8	

TABLE 10.3 data.xlsx ZOU744WS.TXT Nov 14 2022 11: 7
INPUT: 40 PERSON 4 ITEM REPORTED: 40 PERSON 4 ITEM 4 CATS MINISTEP 4.8.2.0

Figure 2. Item Statistics of Misfit Order

Based on the data from Figure 2, the items "Content," "Structure," "Mechanic," and "Diction" exhibit strong fit and reliability with the Rasch model. Here's a detailed description of each item:

1. Content:
   - Outfit MNSQ = 0.70: This value indicates that the item "Content" fits the Rasch model well, being slightly better than the expected value of 1.0 but within the acceptable range (0.5 - 1.5).
   - Outfit ZSTD = 0.23: The value is close to 0, suggesting no significant deviation from the Rasch model.
   - PT Measure Correlation = 0.91: A very high correlation, showing that "Content" is highly consistent with the overall ability measured by the scale.
2. Structure:
   - Outfit MNSQ = 0.61: This value shows that "Structure" fits the Rasch model better than expected, falling below the expected 1.0 and within the acceptable range.
   - Outfit ZSTD = -0.03: This value is very close to 0, indicating no significant deviation.
   - PT Measure Correlation = 0.88: This high correlation indicates that "Structure" contributes well to the measurement.
3. Mechanic:
   - Outfit MNSQ = 0.53: This value indicates that "Mechanic" fits the Rasch model very well, being lower than the expected 1.0 and within the acceptable range.
   - Outfit ZSTD = -0.01: This value is very close to 0, suggesting no significant deviation.
   - PT Measure Correlation = 0.88: A high correlation, showing that "Mechanic" is highly consistent with the overall ability measured by the scale.
4. Diction:
   - Outfit MNSQ = 0.44: This value indicates that "Diction" fits the Rasch model exceptionally well, possibly indicating a slight overfit, but still within the acceptable range.
   - Outfit ZSTD = -0.03: This value is very close to 0, suggesting no significant deviation.
   - PT Measure Correlation = 0.91: A very high correlation, showing that "Diction" is highly consistent with the overall ability measured by the scale.

The data indicates that all items ("Content," "Structure," "Mechanic," "Diction") show an excellent fit with the Rasch model. This is evidenced by their Outfit MNSQ

values being within the acceptable range, Outfit ZSTD values close to 0, and high PT Measure Correlation values. These metrics suggest that these items are reliable and valid measures, contributing significantly to the overall measurement consistency and accuracy of the scale. Selain MNSQ, t-value juga digunakan untuk mengevaluasi misfit yaitu t-value yang signifikan secara statistik (biasanya lebih besar dari ±2) menunjukkan misfit.

In Rasch analysis, the "most misfitting response string" refers to the pattern of a participant's answers that most deviates from the expected Rasch model (Tesio et al, 2024; Rahman et al, 2024). Misfitting response strings can provide insights into participant behavior or items that do not align with the Rasch model assumptions.

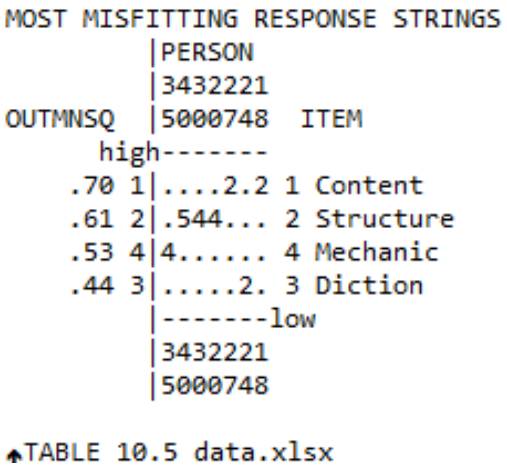

Figure 3. Most Misfitting Response String

In Figure 3, it can be seen that there are four items—content, structure, mechanic, and diction — that sequentially provide the following OUT MNSQ results: "Content" outmnsq = 0.70, "Structure" outmnsq = 0.61, "Mechanic" outmnsq = 0.53, and "Diction" outmnsq = 0.44. Here is the interpretation of the most misfitting response string in Rasch measurement for the item values in Figure 3:

1. Item "Content" (Outfit MNSQ = 0.70) : An Outfit MNSQ value of 0.70 indicates that this item fits the Rasch model reasonably well. While there are some misfitting responses, this value is still within the acceptable range (0.5 - 1.5). Therefore, despite a few discrepancies, this item is considered fairly good.
2. Item "Structure" (Outfit MNSQ = 0.61): An Outfit MNSQ value of 0.61 shows that this item fits the Rasch model better than the "Content" item. There are fewer misfitting responses for this item compared to "Content," and this value is also within the acceptable range. This item can be considered good in terms of fit to the model.
3. Item "Mechanic" (Outfit MNSQ = 0.53): An Outfit MNSQ value of 0.53 indicates that this item fits the Rasch model very well, better than both "Content" and "Structure." There are very few misfitting responses for this item, and this value is within the acceptable range. This item shows an excellent fit to the Rasch model.
4. Item "Diction" (Outfit MNSQ = 0.44): An Outfit MNSQ value of 0.44 indicates that this item is the best fit to the Rasch model among the four items. Although this value is very low and at the lower threshold of the acceptable range (0.5), it shows

that there are almost no misfitting responses. This item can be considered very good, but it should be monitored to ensure that it is not overfitting, which could reduce the uniqueness or variability of responses.

Item "Diction" is the best fit to the Rasch model with the lowest Outfit MNSQ value (0.44), followed by "Mechanic" (0.53), "Structure" (0.61), and "Content" (0.70). Although all items have relatively low Outfit MNSQ values within the acceptable range (0.5 - 1.5), "Diction" shows the best fit to the model, and "Content," while still within the acceptable range, shows the lowest fit among the four items. Overall, all these items demonstrate a reasonable fit to the Rasch model. However, special attention might be needed for the "Diction" item to ensure that overfitting is not occurring. Generally, there are very few misfitting responses for all these items, indicating good reliability in the measurement conducted.

Conclusion

Measuring item fit in Rasch measurement is a crucial process to ensure that the test instrument functions well and provides accurate measurements. This is done using fit statistics such as Infit and Outfit MNSQ. Based on the interpretation of the misfit order Rasch measurement for item fit, all four items ("Content," "Structure," "Mechanic," and "Diction") show a very good fit to the Rasch model. The Outfit Mean Square (MNSQ) values for each item are below the expected value of 1.0 but still within the acceptable tolerance range (0.5 - 1.5), indicating that these items fit the model better than expected. Specifically, the "Content" item has an Outfit MNSQ value of 0.70, the "Structure" item 0.61, the "Mechanic" item 0.53, and the "Diction" item 0.44, with the "Diction" item showing the highest fit.

The Outfit Z-Standard (ZSTD) values for all items are close to 0, indicating no significant deviations from the Rasch model. This means that the responses to these items do not show significant anomalies or mismatches with the model's predictions. The ZSTD values for the "Content" item are 0.23, for the "Structure" item -0.03, for the "Mechanic" item -0.01, and for the "Diction" item -0.03. These values suggest that the items perform as expected within the Rasch model context.

The very high Point Measure Correlation (PT Measure Corr) for all items indicates that they are highly consistent with the overall ability measured by the scale. The correlation values for the "Content" and "Diction" items are 0.91, while for the "Structure" and "Mechanic" items, they are 0.88. These high correlation values show that these items contribute significantly and validly to the measurement of the desired ability. Overall, these four items can be considered very good within the measurement scale used, with the "Diction" item showing the highest fit to the Rasch model.

REFERENCES

Mccreary L. Linda et al. 2013. "Using the Rasch Measurement Model in Psychometric Analysis of the Family Effectiveness Measure." *National Library of Medicine* 62(3):149–59.

Meisel K. et al. 2017. "Subjectivity of Teacher Judgments: Exploring Student Characteristics That Influence Teacher Judgments of Student Ability." *Science Direct* 65:48–60.

Tesio Luigi et al. 2024. "Interpreting Results from Rasch Analysis 2. Advanced Model Applications and the Data-Model Fit Assessment." *Taylor & Francis Online* 46(3):604–17.

Andrich, D. 1978. "A Rating Formulation for Ordered Response Categories."

Psychometrika 43(1):561–573.

Erguvan, Inan Deniz, and Beyza Aksu Dunya. 2020. "Analyzing Rater Severity in a Freshman Composition Course Using Many Facet Rasch Measurement." *Language Testing in Asia* 10(1).

Huang, Hung Yu. 2023. "Modeling Rating Order Effects Under Item Response Theory Models for Rater-Mediated Assessments." *Applied Psychological Measurement* 47(4):312–27.

Jacobs., Holly. L., Stephen, A., Zingkgraf., Deanne. R., Wormuth, V., Faye, H., Jane, B., Hughey. 1981. *Testing ESL Composition: A Practical Approach*. Rowley: Newbury House Publishers, Inc.

Li, Guangming, Yuxi Pan, and Weijun Wang. 2021. "Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions." 12(July):1–10.

Linacre, J. M. (. 2002. "What Do Infit and Outfit Mean-Square and Standardized Mean?" *Rasch Measurement Transaction* 16:878.

Misbach, I. H., & Sumintono, B. 2014. "Pengembangan Dan Validasi Instrumen 'Persepsi Siswa Tehadap Karakter Moral Guru' Di Indonesia Dengan Model Rasch." . . *PROCEEDING Seminar Nasional Psikometri* 148–162.

Rahman, Yenni Arif. 2023. "Person and Item Validity and Reliability in Essay Writing Using Rasch Model." 15(1):41–55.

Rahman, Yenni Arif, Nurhayati S., Fiza Asri Fauziah Habibah, and Fadilah Fadilah. 2024. *Holistic Rubric Validity and Reliability in Essay Assessment Using Rasch Model in Blended Learning Program*. Atlantis Press SARL.

Reise, S. .. 1990. "A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT." *Applied Psychological Measurement* 14(2):127–37.

Rost, J., & von Davier, M. 1994. "A Conditional Item-Fit Index for Rasch Model." *Applied Psychological Measurement* 18(2):171–82.

Stemler, Steven E., and Adam Naples. 2021. "Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line." *Practical Assessment, Research and Evaluation* 26(May):1–16.

Sumintono, B. & Widhiarso, W. 2013. *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial*. Trim Komunikata Publishing House.

Tan, S. 2013. "Validation of an Analytic Rating Scale for Writing: A Rasch Modeling Approach." *Tabaran Institute of Higher Education. Iranian Journal of Language Testing* 3(1).

Uto, Masaki. 2022. "A Bayesian Many-Facet Rasch Model with Markov Modeling for Rater Severity Drift." *Behavior Research Methods* (October).