Lecture Notes on Language and Literature (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2025.080502
ISSN 2523-5869 Vol. 8 Num. 5

Evaluating the Psychometric Quality of Classroom English Assessments through the Rasch Model

Jiayun Xue[1,a], Meihua Chen[1,b,*], Jiayi Su[2,c]

[1]*School of Foreign Languages, Southeast University, No. 2, Southeast University Road, Jiangning District, Nanjing, 211189, China*
[2]*Wujiang Yunlong Experimental School Affiliated to Shanghai WFL Education, No. 555, Huxin West Road, Wujiang District, Suzhou, Jiangsu, 215299, China*
[a]*stephaniexue2002@foxmail.com*, [b]*meihuachen123@126.com*, [c]*1643650771@qq.com*
[*]*Corresponding author*

Keywords: Rasch Model; English Test; Middle School English; Test Quality Analysis

Abstract: This study applied Rasch modeling to evaluate the psychometric quality of a regional middle school English assessment. Data were drawn from a stratified cluster sample of 598 seventh-grade students across seven schools. Analyses included item fit statistics, separation indices, a Wright map to examine how well the test measured student ability and distinguished proficiency levels. Results showed strong reliability and good model-data fit, with most Infit and Outfit MNSQ values within acceptable ranges. The test was well targeted for average students but contained few very difficult or easy items, limiting precision at the extremes. Content analysis also revealed redundancy in items testing similar vocabulary and grammar. To improve measurement efficiency and fairness, the study recommends adding both challenging and easier items and refining overlapping content. The findings demonstrate the value of Rasch analysis in guiding evidence-based improvements to classroom-based language assessments.

1. Introduction

Language assessment is a hybrid discipline integrating applied linguistics and measurement expertise. Applied linguistics informs conceptions of language ability, while measurement ensures assessments are reliable and valid. Such knowledge is central to language assessment literacy and language testing [1]. Among measurement approaches, Rasch modeling has been widely used since the 1980s [2] to evaluate test quality and improve validity and fairness [3].

Rasch analysis enhances test accuracy and fairness in achievement, placement, and proficiency tests [4]. Yet, despite its broad use in standardized testing, Rasch modeling is seldom applied systematically to classroom-based summative assessments [5]. Existing research focuses largely on high-stakes exams at high school level and above, such as college entrance tests [6] or professional certifications [7], with little attention to lower secondary education, especially final English exams for younger learners.

To address this gap, this study uses the Rasch model to evaluate a seventh-grade final English exam, focusing on item fit, reliability, separation indices, and item-person alignment. The aim is to

demonstrate how Rasch analysis can improve classroom assessments, bolster validity and fairness, and support more equitable evaluation of young learners' language proficiency.

2. Literature review

Language assessment is vital for evaluating proficiency and guiding instruction, with quality defined by reliability, validity, and fairness. Validity concerns whether scores reflect the intended construct and support appropriate interpretations. Reliability refers to score consistency across instruments, raters, and occasions, while fairness ensures equitable interpretation for all test-takers [8]. Although large-scale tests undergo rigorous analysis, classroom assessments often lack systematic evaluation, raising concerns about their validity and fairness.

The Rasch model, a family of probabilistic measurement models, offers a framework for analyzing language assessments. It assumes unidimensionality and maps persons and items onto a common logit scale. Fit statistics such as Infit and Outfit MNSQ help identify misfitting items. Since its adoption in language testing [2], the model has been used to analyze item difficulty, person ability, and model-data fit [3]. Multi-faceted Rasch measurement (MFRM) further allows analysis of rater severity and task difficulty, serving as a microscope for rating patterns.

Despite widespread use in high-stakes testing, Rasch analysis is rarely applied to classroom-based, medium- or low-stakes exams, especially at the lower secondary level. While exams such as seventh-grade English finals influence instruction and placement, they remain understudied [1]. An exception is Zhan & Bai (2024) [9], who applied Rasch analysis to eighth-grade science tests, yet English assessments at this level are still overlooked.

To address this gap, this study employs Rasch analysis to evaluate a seventh-grade final English exam, focusing on:

(1) What is the overall psychometric quality of the exam?

(2) How well do item difficulties align with student abilities?

(3) Which items show misfit, and what are the implications for validity and diagnostic usefulness?

3. Methodology

3.1 Participants

The study analyzed the complete test records of 598 seventh-grade students, selected via stratified cluster sampling from a regional unified final English exam.

3.2 Instrument

The 55-item multiple-choice exam assessed listening and reading comprehension, along with basic language knowledge. Items were dichotomously scored (0/1), except for reading comprehension which used partial credit (0/2) to capture partial understanding.

3.3 Procedure

Answer sheets were collected and raw scores were entered into Excel. Data were analyzed using Winsteps 3.72.3 to estimate item difficulty, person ability, reliability, and fit statistics (Infit/Outfit MNSQ). Visual outputs (e.g., Wright maps) were used to examine measurement characteristics and item-person alignmen. The proceudre was displayed in Figure 1.

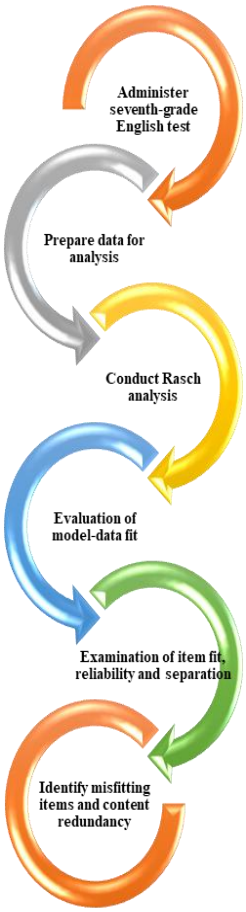

Figure 1: Research procedure

3.4 Data Analysis

The Rasch model was applied to measure latent ability based on response patterns. The analysis evaluated unidimensionality, item-person fit, and the alignment between student ability and item difficulty on a common logit scale, providing evidence for test quality and refinement.

4. Results

4.1 The overall quality of the exam paper

In Rasch analysis, to evaluate the overall quality of the test, software such as Winsteps 3.72.3 is typically used to calculate key indicators, including measure (average ability and item difficulty), separation (the test's capacity to distinguish between different ability levels), reliability (internal consistency), and fit statistics such as Infit MNSQ, Outfit MNSQ, and their standardized forms (ZSTD). As shown in Figure 2, the average person measure is -0.12, suggesting that overall test difficulty is reasonably aligned with the students' ability levels. The person separation index is 3.43 (>2), indicating that the test can effectively differentiate between students of varying proficiency. The person reliability is 0.92 (>0.7), reflecting high internal consistency. At the item level, the average item measure is close to 0.00, the item separation index is as high as 11.59, and item reliability reaches 0.99, all pointing to stable and reliable estimates of item difficulty. Regarding fit, the Infit MNSQ values for both persons and items are around 1.03, which falls within the acceptable

range of 0.5 to 1.5, demonstrating good model and data fit. Although the item ZSTD mean is slightly beyond the recommended range (-2.4), this suggests only mild overfit, meaning the responses were slightly more predictable than expected, which is generally not harmful. Overall, the analysis shows that the test has good targeting, high reliability and separation, and acceptable model fit, making it suitable for further Rasch-based evaluation.

4.2 Unidimensionality test of the exam paper

The Rasch model, as a single-parameter item response theory (IRT) model, fundamentally relies on the assumption of unidimensionality. This means that students' performance on the test should primarily reflect a single underlying latent trait, that is, their English language proficiency. Other factors, such as guessing strategies, test anxiety, or unrelated cognitive skills, should have minimal influence. Testing for unidimensionality is therefore an essential diagnostic step in Rasch analysis, because if this assumption does not hold, any further estimates of item difficulty, person ability, or fit statistics may be biased or misleading.

To evaluate unidimensionality in this study, the researcher used the standardized residual contrast plot generated by Winsteps software. Methodologically, after fitting the Rasch model to the data, Winsteps calculates residuals, that is, the differences between the actual observed responses and the expected responses predicted by the model. Then, a principal components analysis (PCA) of these residuals is conducted to identify whether there are any substantial secondary dimensions (contrasts) that explain leftover variance not captured by the Rasch dimension.

In this plot, the horizontal axis represents item difficulty measures (indicating how challenging each item is for the sample), and the vertical axis shows the standardized residual contrasts, which reflect potential correlations with secondary traits or unintended dimensions. Each letter (A, B, C, etc.) corresponds to a single test item. According to widely accepted criteria, if most residual contrast values fall between -0.4 and +0.4, this suggests items are sufficiently related to the primary latent trait, supporting unidimensionality.

From Figure 3, it can be observed that the majority of items cluster within the -0.4 to +0.4 interval. This finding suggests that the test is largely measuring one common construct, students' overall English proficiency, and other unintended factors have minimal systematic influence. This supports the validity of applying the Rasch model to further analyze item difficulty, person ability, reliability, and fit statistics.

Nevertheless, there is one item labeled x whose residual contrast values slightly exceeds this recommended range. Methodologically, this means this item may load onto an additional dimension, perhaps reflecting separate skills such as reading comprehension subskills, vocabulary knowledge, or test-taking strategies. Although this does not necessarily invalidate the overall test, it indicates areas where test developers may consider conducting a qualitative content review or further statistical checks to understand why these items behave differently.

In conclusion, the methodological process of residual PCA and contrast plotting provides empirical evidence that the test shows acceptable unidimensionality. This justifies the use of the Rasch model in subsequent stages of test analysis, ensuring that estimates of item difficulty and person ability are valid and meaningful within the intended construct of English proficiency.

4.3 The wright map of the exam paper

The Rasch model places both student ability and item difficulty on the same linear logit scale, shown visually in the Wright map (Figure 4), so that we can directly compare their distributions. In this map, the vertical dashed line represents the shared logit scale: higher values toward the top indicate higher ability for students or greater difficulty for items.

On the left side of the scale, the distribution of students' abilities is marked by "#", each symbol representing a number of students. On the right side, individual items are listed according to their estimated difficulty measures. The letters "M", "S", and "T" indicate the mean (M), one standard deviation from the mean (S), and two standard deviations from the mean (T), respectively. Moving from the bottom to the top of the scale, logit values increase, reflecting higher student ability and greater item difficulty.

From Figure 3, we can see that most students' ability measures are concentrated between 0 and +2 logits, forming a negatively skewed distribution. This suggests that, overall, the test was relatively easy for this sample: most students performed around or above the average item difficulty.

However, the map shows that while many items cluster between 0 and +1 logits (matching the bulk of the students' abilities), there are few items located above +2 logits. This indicates a lack of very difficult items capable of effectively differentiating among the higher-ability students. As a result, the test may not be sufficiently challenging for the most advanced students in the sample.

Methodologically, the Wright map helps visualize whether the test targets the intended population. Ideally, items should cover the full range of student abilities. Gaps in item distribution (such as the absence of items >+2 logits) indicate areas where test developers could introduce harder questions to better assess high performers. Clusters of items with very similar difficulty, such as many around 0 logits, suggest potential redundancy and an opportunity to revise or diversify item difficulty.

Overall, this analysis confirms that while the test aligns well with most students' abilities, it may lack enough high-difficulty items to fully measure and differentiate the top-performing students. Such insights from the Wright map are critical for guiding test improvement and ensuring balanced measurement across the full ability spectrum.

4.4 Summary of the measured person and item

As in Figure 5 and 6, Rasch analysis of the 7th-grade English final exam indicates the test is generally well-targeted for the majority of students, with overall good reliability. However, it lacks sufficient high-difficulty items to differentiate top-performing students and shows limited informativeness for the lowest-ability students.

Key findings include a well-matched average item difficulty and student ability, high item reliability (0.99), and a person reliability of 0.92–0.94, supporting the separation of students into about five proficiency levels. Nevertheless, the test's discriminative capacity is constrained at the ability extremes due to a narrow item difficulty range and some content redundancy.

Recommendations involve adding more challenging items (above 2 logits), incorporating easier items to reduce measurement error, and removing redundant questions to improve overall measurement precision and coverage.

5. Discussion

Rasch analysis indicates that this final English exam effectively measures general proficiency among seventh graders, as shown by high reliability and a range of item difficulties. The test reliably distinguishes students across approximately five proficiency levels [10], and unidimensionality was confirmed, supporting that it primarily reflects a single underlying trait [1].

However, the concentration of items around medium difficulty limits accurate measurement at the ability extremes, echoing findings from large-scale tests where balanced difficulty distributions are essential [11]. The presence of overfitting and underfitting items—potentially due to ambiguous phrasing or content familiarity—along with vocabulary and structural redundancy, may reduce diagnostic precision, consistent with prior classroom assessment studies [1].

In summary, while the exam demonstrates solid psychometric quality for typical learners, it would benefit from adding more challenging items, reducing redundancy, and refining misfitting items to better capture the full spectrum of student abilities and support differentiated instruction.

6. Conclusion and implications

This Rasch analysis confirms the overall soundness of the seventh-grade English exam, showing good person-item targeting and high reliability. However, the absence of extreme-difficulty items limited discrimination at the ability extremes, and several items showed overfit or underfit. To enhance the test, we recommend introducing more challenging and easier items, reducing content redundancy, and complementing quantitative analysis with qualitative methods like expert review or think-aloud protocols. Ultimately, applying Rasch modeling regularly can make classroom assessments more valid, fair, and instructionally useful.

PERSON	598 INPUT	598 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	52.9	55.0	-.12	.27	1.03	-.2	1.03	-.2
S.D.	16.7	.0	1.02	.08	.48	2.8	.46	2.5
REAL RMSE	.28	TRUE SD	.98	SEPARATION	3.43	PERSON RELIABILITY		.92

ITEM	55 INPUT	55 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	575.2	598.0	.00	.09	1.03	-2.4	1.03	-2.1
S.D.	191.5	.0	1.03	.02	.82	9.3	.73	8.9
REAL RMSE	.09	TRUE SD	1.02	SEPARATION	11.59	ITEM RELIABILITY		.99

Figure 2: Overall quality plot

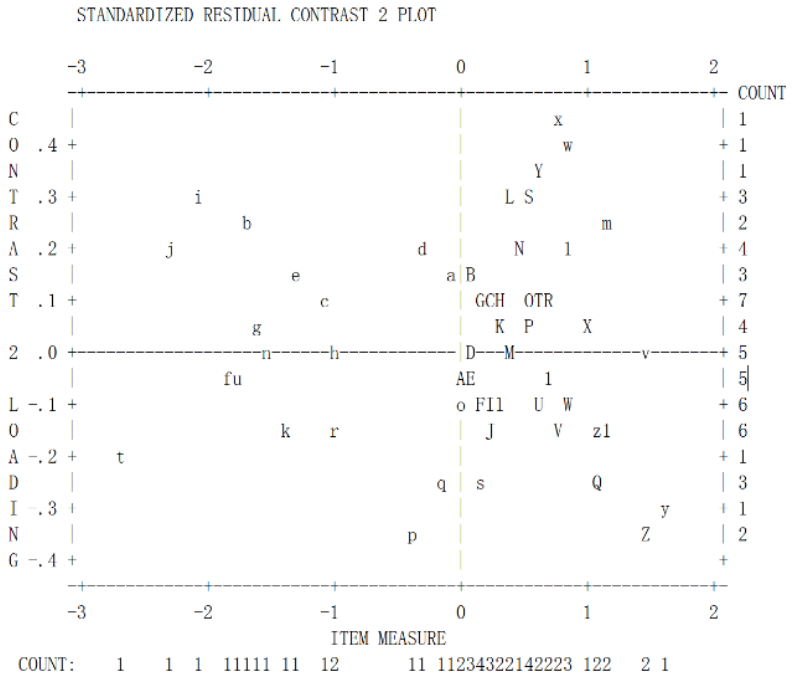

Figure 3: Standardized residual plot

```
---------------------------------------------------------------------------
----

          PERSON - MAP - ITEM
              <more>|<rare>
   2                +T
                   T|
                    |  Q25
                    |  Q28    Q49
                    |
            .#####  |  Q22    Q48
   1 .############  +S Q19    Q20    Q30
         .####### S|  Q23    Q26    Q50    Q51    Q6
      ############  |  Q1     Q21    Q3     Q52
         .########  |  Q17    Q18    Q2     Q27    Q29    Q4
          .######   |  Q15    Q24    Q53    Q7
         .########  |  Q10    Q11    Q12    Q13    Q14    Q16    Q45    Q8
                    |  Q9
   0         #####  +M Q46    Q5
            .##### M|  Q47    Q55
           .######  |  Q40    Q54
           ######   |
             .###   |
           .#####   |
  -1       .#####  +S Q34    Q36
             .### S|  Q37
            .####   |  Q42    Q44
                .   |  Q38    Q41
              .##   |  Q43
              .##   |  Q35    Q39
  -2           .## T+T Q33
               .#   |
               ##   |  Q32
                .   |
                #   |  Q31
                    |
  -3                +
                .   |
                    |
                    |
                    |
                    |
  -4                +
                    |
                    |
                    |
                    |
                    |
  -5                +
              <less>|<frequ>
EACH "#" IS 5. EACH "." IS 1 TO 4
```

Figure 4: The wright map

SUMMARY OF 598 MEASURED (EXTREME AND NON-EXTREME) PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	52.9	55.0	-.12	.25				
S.D.	16.7	.0	1.02	.07				
MAX.	75.0	55.0	1.16	1.83				
MIN.	.0	55.0	-7.12	.23	.27	-5.9	.27	-5.4
REAL RMSE	.28	TRUE SD	.98	SEPARATION	3.43	PERSON RELIABILITY		.92
MODEL RMSE	.26	TRUE SD	.98	SEPARATION	3.84	PERSON RELIABILITY		.94
S.E. OF PERSON MEAN = .04								

Figure 5: Summary of 598 measured person

SUMMARY OF 55 MEASURED (NON-EXTREME) ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	575.2	598.0	.00	.07	1.03	-2.4	1.03	-2.1
S.D.	191.5	.0	1.03	.00	.82	9.3	.73	8.9
MAX.	1032.0	598.0	1.59	.10	2.48	9.9	2.47	9.9
MIN.	284.0	598.0	-2.70	.07	.27	-9.9	.32	-9.9
REAL RMSE	.09	TRUE SD	1.02	SEPARATION	11.59	ITEM RELIABILITY		.99
MODEL RMSE	.07	TRUE SD	1.03	SEPARATION	13.82	ITEM RELIABILITY		.99
S.E. OF ITEM MEAN = .14								

UMEAN=.0000 USCALE=1.0000
ITEM RAW SCORE-TO-MEASURE CORRELATION = -1.00
32835 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 52459.30 with 32183 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .5555

Figure 6: Summary of 598 measured item

References

[1] Effatpanah, F., Baghaei, P., Ravand, H., & Kunina-Habenicht, O. (2024). Fitting the mixed Rasch model to the listening comprehension section of the IELTS: Identifying latent class differential item functioning. International Journal of Testing, 25(1), 50–89.

[2] Baker, R. (1987). An investigation of the Rasch model in its application to foreign language proficiency testing. Unpublished PhD thesis. University of Edinburgh, Edinburgh.

[3] McNamara, T., Knoch, U., & Fan, J. (2019). Fairness, justice and language assessment. Oxford: Oxford University Press.

[4] Dunya, B. A., & Wind, S. A. (2025). Exploring the effects of small item pools on examinee achievement estimates for computer-adaptive tests: A simulation study. International Journal of Testing, 25(2), 127–143.

[5] Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. Language Testing, 31(4), 501–527.

[6] Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. Assessing Writing, 32(2), 43–56.

[7] Isbell, D. R. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. Assessing Writing, 34(1), 37–49.

[8] Kang, O., Yan, X., Kostromitina, M., Thomson, R., & Isaacs, T. (2024). Fairness of using different English accents: The effect of shared L1s in listening tasks of the Duolingo English test. Language Testing, 41(2), 263–289.

[9] Zhan, L., & Bai, Y. (2024). Assessment of middle school students' scientific literacy based on the Rasch model. Journal of Southeast University, 26(2), 45–48.

[10] Aryadoust, V., & Luo, L. (2022). The typology of second language listening constructs: A system atic review. Language Testing, 40, 375–409.

[11] Chen, Y., & Zhou, R. (2018). An investigation of the quality of English perfect fill-in-the-blank questions in the college entrance examination based on Rasch model. Foreign Language Testing and Teaching, (1), 39–47, 64.