Scientific and Social Research, 2025, Volume 7, Issue 11
https://ojs.bbwpublisher.com/index.php/SSR
Online ISSN: 2981-9946
Print ISSN: 2661-4332

A Many-Facet Rasch Model Analysis of Rater-Criterion Interaction in the Assessment of Student Post-Machine Translation Editing Competence

Chun-guang Tian*

School of Foreign Languages, Shandong University of Aeronautics, Binzhou 256600, China

**Author to whom correspondence should be addressed.*

Abstract: With the advancement of machine translation, post-editing (PE) has become a dominant workflow, making the accurate assessment of post-machine translation editing competence (PMTE) critical. However, performance-based PMTE assessments are vulnerable to subjective rater effects, compromising their validity. This study employs the Many-Facet Rasch model (MFRM) to conduct an in-depth analysis of rater-criterion interaction, a complex bias in PMTE evaluation. The research involved 144 examinees translating a scientific text and four expert raters assessing the outputs using four criteria: Logical relations, completeness, terminology, and fluency. The MFRM analysis successfully calibrated examinee ability, demonstrating high reliability (.86), and revealed significant variations in rater severity and criterion difficulty. Critically, the analysis identified specific quality control issues, including inconsistent scoring by one rater and ambiguity in the "Terminology" criterion. The bias analysis uncovered significant rater-criterion interactions. These findings demonstrate that MFRM is a powerful diagnostic tool that transforms assessment from a purely evaluative act into a mechanism for data-driven improvement. It provides objective, actionable evidence for refining scoring rubrics and conducting targeted rater training, thereby enhancing the fairness and validity of PMTE assessment.

Keywords: Post-machine translation editing competence (PMTE); Many-Facet Rasch model (MFRM); Rater-Criterion interaction; Translation quality assessment; Rater effects

Online publication: December 12, 2025

1. Introduction

With the rapid advancement of artificial intelligence technology, the quality of machine translation (MT) has significantly improved, leading to post-editing (PE) gradually becoming one of the dominant working modes in the translation industry [1]. In this new workflow, the role of the translator has shifted from a traditional "translator"

to a "post-editor", and their core competence has correspondingly expanded from pure translation skills to post-machine translation editing competence (PMTE). PMTE not only requires translators to possess solid bilingual and bicultural knowledge but also demands comprehensive skills in quickly identifying MT errors, efficiently correcting the translation, and assessing translation quality [2]. Consequently, the accurate assessment of students' PMTE competence has become a critical issue in translation education and professional practice.

The assessment of PMTE competence typically takes the form of a performance assessment, where students are required to edit the output of a machine translation system, and human raters then evaluate the results based on a set of predefined scoring criteria. However, this subjective scoring process is highly susceptible to Rater effects, such as differences in rater severity or leniency, and inconsistencies in how raters interpret and apply specific scoring criteria [3]. These factors introduce measurement error, which distorts the estimation of students' true PMTE ability and severely compromises the reliability and validity of the assessment results.

To address the psychometric challenges inherent in performance assessments, this study proposes the use of the Many-Facet Rasch Model (MFRM) for an in-depth analysis of student PMTE competence scoring data. As an extension of the Item Response Theory (IRT) model, MFRM can separate and place multiple "facets" onto the same linear scale for objective measurement [4]. The primary focus of this study is to leverage the powerful capabilities of MFRM to separate and quantify the Rater-Criterion Interaction, thereby identifying systematic biases in how raters apply specific scoring criteria. By doing so, this research aims to provide a more scientific and objective method for PMTE competence assessment and offer empirical evidence for optimizing PMTE scoring criteria and rater training.

2. Literature review

2.1. Definition and assessment of post-machine translation editing competence (PMTE)

PMTE is an evolution and extension of translation competence in the era of machine translation [5]. Scholars generally agree that PMTE is a composite competence that includes not only traditional translation competence elements (such as linguistic and cultural competence) but also specifically emphasizes the technical and cognitive skills necessary in the MT environment [2]. Specifically, PMTE involves multiple dimensions, including error identification, error classification, correction efficiency, and quality control of the MT output [6].

In the practice of PMTE assessment, quality assessment frameworks form the basis for defining scoring criteria. Currently, the widely adopted frameworks in both industry and academia primarily include the Multidimensional Quality Metrics (MQM) and the Dynamic Quality Framework (DQF) [1]. MQM provides a detailed, hierarchical system for classifying error types, covering seven main dimensions such as accuracy, fluency, terminology, and style, and assigns different severity levels to each error type [7]. DQF, on the other hand, focuses on integrating quality assessment with the translation process (e.g., post-editing efficiency) [8]. These detailed scoring criteria (such as "terminology errors" or "grammatical errors" in MQM) constitute the assessment items or scoring dimensions in PMTE performance tests. However, when human raters use these detailed criteria, the objectivity of the scoring results is challenged by differences in rater interpretation of the criteria and subjective judgment [3].

2.2. Application of the Many-Facet Rasch model (MFRM) in performance assessment

The Rasch model is a crucial measurement model in psychometrics, whose core advantage lies in enabling

objective measurement. It can independently estimate examinee ability and item difficulty on the same scale, and these estimates are independent of sample or item characteristics [4]. The Many-Facet Rasch model (MFRM), an extension of the Rasch model proposed by Linacre in 1989, is designed to incorporate other influencing factors in performance assessment—beyond examinee ability and item difficulty—such as raters, scoring criteria, and testing occasions, as independent "facets" for simultaneous estimation [9]. The basic MFRM equation converts the parameters of all facets (ability, difficulty, severity, etc.) into Log-odds Units (Logits), allowing for direct comparison between different facets [10].

MFRM has proven to be a powerful tool in the field of language and translation assessment. For instance, in speaking and writing tests, MFRM is widely used to detect and calibrate rater severity and leniency [5]. Through MFRM analysis, researchers can identify raters who are overly harsh or lenient and statistically correct their scores to obtain a fairer estimate of examinee ability [11]. In the field of translation testing, the study by Tseng et al. applied MFRM to validate the reliability and validity of translation test items and successfully separated and modeled rater severity, confirming the model's effectiveness in translation competence assessment [12]. These studies provide a solid theoretical and methodological foundation for applying MFRM to PMTE competence assessment in this research.

2.3. Analysis of Rater-Criterion interaction: Research gap and necessity

In performance assessments, besides the overall differences in rater severity, a more complex and subtle issue is the Rater-Criterion Interaction [6]. This interaction means that a rater may not be overly harsh overall but exhibits systematic severity or leniency toward a specific scoring criterion (e.g., "terminology errors" or "fluency" in PMTE assessment) [7]. This bias leads to inconsistencies in the rater's scoring behavior across different dimensions, thereby distorting the student's true performance across various ability dimensions.

One of the key advantages of MFRM is its ability to model and quantify this complex interaction [8]. By analyzing the fit statistics of the interaction term, researchers can identify raters who behave unusually when applying specific scoring criteria and quantify the extent of this bias [6]. For example, Eskin used MFRM to analyze the interaction between raters and scoring criteria in writing assessment, finding significant differences in how raters applied different criteria, thus emphasizing the importance of separating and correcting this interaction [9].

However, despite the widespread application of MFRM in language assessment and the use of highly detailed scoring criteria like MQM/DQF in PMTE assessment, current research has not fully utilized MFRM to specifically investigate the interaction between raters and these detailed scoring criteria in PMTE competence assessment. Traditional MFRM applications mostly focus on separating the main effects of examinee ability, item difficulty, and rater severity [13]. This research will fill this gap by using MFRM's interaction analysis to provide a new perspective and empirical evidence for the objective measurement of PMTE competence.

3. Research design

3.1. Research context and objectives

The primary objective was to develop and validate a robust measurement model for assessing the quality of English-to-Chinese scientific and technical translations. The specific research objectives were:

(1) To quantitatively estimate the translation ability of examinees on a shared interval scale, adjusted for rater and criterion effects.

(2) To evaluate the consistency and severity of the raters, identifying any significant deviations in their scoring patterns.
(3) To calibrate the difficulty levels of the predefined assessment criteria to understand their relative contribution to the overall quality score.

3.2. Participants

The study involved two distinct groups of participants:

Examinees (Translators): A total of 144 participants were involved in the translation task. These individuals were students or professionals with training in scientific and technical translation. To ensure anonymity and objectivity in the analysis, each examinee was assigned a numerical code.

Raters (Experts): Four expert raters (Rater ID: R1, R2, R3, R4) were recruited to evaluate the translations. All raters were certified translators or subject-matter experts with extensive experience (minimum of 5 years) in the field of scientific and technical translation between English and Chinese.

3.3. Instruments and materials

Translation task: All examinees were tasked with translating a standardized English source text of approximately 500 words into Chinese. The text was selected from a peer-reviewed journal in the field of biotechnology, ensuring it contained domain-specific terminology, complex logical structures, and stylistic conventions representative of scientific discourse.

Assessment criteria: Translation quality was evaluated along four pre-defined, independent criteria. These criteria were adapted from House's model of translation quality assessment, which provides a systematic framework for analyzing a translation against its source text based on linguistic and functional equivalence [13]. The four criteria operationalized for this study are:

Logical relations (LR): Corresponding to House's textual and pragmatic dimensions, this criterion assesses the accuracy and coherence of logical connectors, cause-effect relationships, and the overall argumentative flow of the translated text.

Completeness (CP): Reflecting the House's concept of textual equivalence, this criterion measures the degree to which all informational content from the source text was rendered in the target text without omission or addition.

Terminology (TM): Aligned with the lexical and semantic dimensions of House's model, this criterion evaluates the precise and consistent translation of domain-specific technical terms and concepts.

Fluency (FL): This criterion, related to House's grammatical and stylistic analysis, assesses the linguistic quality of the target text, including grammatical correctness, naturalness of expression, and adherence to target language conventions.

Rating scale: A 5-point Likert-type scale was used for each criterion, ranging from 1 (Unacceptable) to 5 (Excellent). Raters provided a single integer score for each of the four criteria for every translation.

3.4. Procedure

The data collection was conducted in a controlled environment. First, all examinees completed the translation task within a specified time limit. Subsequently, the 144 translated target texts were randomized and distributed to the four raters. Raters evaluated the translations independently over a two-week period. To mitigate potential

order effects, the sequence in which each rater received the translations was randomized. Raters were blind to the identities of the examinees. Prior to the formal rating process, all raters participated in a calibration session to ensure a shared understanding of the four assessment criteria and the scoring rubric, thereby enhancing inter-rater reliability.

3.5. Data analysis

To account for the multi-faceted nature of the data, a Many-Facet Rasch measurement (MFRM) model was employed using Facets 64-bit software (version 4.3.1). This psychometric approach is superior to classical test theory as it simultaneously estimates parameters for examinee ability, rater severity, and criterion difficulty on a common logit scale, while providing detailed diagnostic information about model fit and data quality.

4. Results

4.1. Examinee ability estimates

The MFRM analysis successfully estimated the translation ability for all 144 examinees on a linear logit scale, adjusted for the severity of raters and the difficulty of criteria. The examinee ability measures ranged from a maximum of 7.89 logits for examinees who received perfect scores to a minimum of -3.75 logits. The mean ability of the examinee cohort was 2.35 logits (Sample SD = 2.53).

The reliability of the examinee measures was .86, with a separation of 2.44, indicating that the assessment could distinguish approximately 3.59 statistically distinct strata of examinee ability. A fixed chi-squared test confirmed significant variability in examinee abilities (x^2 = 978.3, df = 143, P < .001).

While most examinees demonstrated acceptable fit, one examinee (Examinee 31) exhibited misfit, with an Infit MnSq of 2.28 (ZStd = 2.6), suggesting an unpredictable pattern of scores. A total of 40 observations (5.0%) were in the extreme score categories (all 1s or all 5s).

4.2. Rater severity and consistency

The analysis revealed significant differences in rater severity. Rater 4 (R4) was the most lenient with a severity measure of 1.02 logits, followed by Rater 2 (R2) at 0.94 logits. Rater 1 (R1) and Rater 3 (R3) were the most severe, both with a measure of -0.98 logits. These differences were statistically significant (x^2 = 238.7, df = 3, P < .001). The rater separation was 7.63, with a reliability of .98, indicating a high degree of distinction among the raters' severity levels.

Regarding consistency, three raters (R1, R3, R4) demonstrated acceptable fit, with Infit MnSq values ranging from 0.83 to 0.89. However, Rater 2 (R2) showed a lack of consistency, with an Infit MnSq of 1.41 (ZStd = 3.4), which exceeds the conventional upper limit of 1.3. The overall exact agreement between raters was 36.8%, which was close to the expected agreement of 39.4%.

4.3. Criterion difficulty calibration

The four assessment criteria were calibrated on the same logit scale, revealing significant differences in their difficulty. "Completeness" was the easiest criterion (Measure = 1.36 logits), while "Fluency" was the most difficult (Measure = -0.57 logits). The other two criteria, "Terminology" (Measure = -0.26 logits) and "Logical Relations" (Measure = -0.52 logits), fell in between. These differences were statistically significant (x^2 = 141.7, df = 3, P < .001). The criterion separation was 6.13, with a reliability of .97, indicating that the criteria were well-

differentiated in terms of difficulty.

With respect to the functioning of the criteria, "Terminology" exhibited a higher-than-expected Infit MnSq of 1.29 (ZStd = 2.6), suggesting potential ambiguity in how this criterion was interpreted or applied across different translations. The other three criteria showed good fit to the model (Infit MnSq from 0.81 to 1.02).

5. Discussion

This discussion interprets the results of the Many-Facet Rasch measurement (MFRM) analysis, elucidating the implications of each key finding for the validity and reliability of the translation quality assessment.

The analysis successfully demonstrated that the assessment instrument possesses strong construct validity in measuring translation ability. The high reliability coefficient (.86) and substantial separation index (2.44) confirm that the test can reliably differentiate among multiple strata of examinee proficiency. This indicates that the items (i.e., the translation task and criteria) are working together coherently to measure a single underlying construct. The identification of one misfitting examinee does not undermine the test's validity but rather showcases the diagnostic power of the Rasch model in flagging atypical response patterns for further review.

A primary contribution of the MFRM analysis is its ability to deconstruct and adjust for rater effects. The results revealed statistically significant differences in rater severity, with R4 being the most lenient and R1/R3 the most severe. Without this statistical adjustment, raw scores would be confounded by these rater effects, making fair comparisons between examinees impossible. The MFRM model successfully partitioned this variance, producing calibrated "fair scores." More critically, the analysis identified a specific quality control issue with Rater 2. This is distinct from severity; it indicates a lack of adherence to the measurement model, suggesting a need for targeted feedback or retraining to align their judgments with the rubric.

The calibration of assessment criteria provided empirical evidence for the construct's internal structure. The finding that "Completeness" was the easiest criterion and "Fluency" the most difficult aligns with the cognitive demands of evaluation; checking for informational omission is a more objective task than judging the subtle nuances of linguistic naturalness. This logical hierarchy of difficulty supports the content validity of the chosen criteria. However, the misfit of the "Terminology" criterion (Infit MnSq = 1.29) is a significant finding. It suggests that this criterion was not applied consistently by raters across different texts, potentially due to varying levels of terminological density or ambiguity in the source texts. This points directly to a weakness in the measurement instrument itself, specifically a need to refine the rubric to provide clearer, more operationalized guidelines for scoring terminology.

6. Conclusion

This study successfully applied the Many-Facet Rasch measurement (MFRM) model to develop and validate a framework for assessing English-to-Chinese scientific and technical translations. The results provide robust empirical evidence for the model's utility and offer significant implications for translation assessment practice.

The primary findings and their implications are as follows. First, the results confirmed that the assessment instrument, when analyzed with the MFRM model, is highly reliable and capable of validly distinguishing among different levels of translation ability. Second, and most importantly, the study revealed the immense power of MFRM as a diagnostic tool. It moved beyond raw scores to precisely quantify and isolate the effects of different facets. The analysis identified significant variations in rater severity, pinpointed specific inconsistencies in one

rater's scoring behavior, and detected ambiguity in the application of the "Terminology" criterion. The key implication is that MFRM transforms assessment from a purely evaluative act into a mechanism for systematic, data-driven quality improvement. It provides objective, actionable evidence for targeted rater training and rubric refinement, thereby enhancing the fairness and validity of the entire assessment system.

In conclusion, this study not only validates a scientific model for translation quality assessment but, more importantly, demonstrates how psychometric tools can be used to drive the refinement and professionalization of assessment practices in the field of translation studies.

Disclosure statement

The author declares no conflict of interest.

References

[1] Lommel A, Melby AK, 2018, MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). In Proceedings of the 21st Annual Conference of the European Association for Machine Translation, 21–30.

[2] Robert IS, Schrijver I, Ureel JJ, 2022, Measuring Translation Revision Competence and Post-editing Competence in Translation Trainees: Methodological Issues. The Interpreter and Translator Trainer, 16(3): 329–349.

[10] Myford CM, Wolfe EW, 2003, Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. Journal of Applied Measurement, 4(4): 386–422.

[4] Boone WJ, 2016, Rasch Analysis for Instrument Development: Why, When, and How. CBE—Life Sciences Education, 15(4): ar54.

[5] Nitzke J, 2019, Risk Management and Post-editing Competence. The Journal of Specialised Translation, 2019(31): 126–146.

[6] Myford CM, Wolfe EW, 2004, Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. Journal of Applied Measurement, 5(2): 189–227.

[9] Eskin D, 2023, A Many-Facets Rasch Analysis of Facet Main Effects and Interactions in a Writing Assessment. Journal of Applied Measurement, 24(1): 1–16.

[8] Dismukes RK, 2000, Multifacet Rasch Analysis of Rater Training Effects on Scoring Performance. American Institutes for Research.

[9] Linacre JM, 1989, Many-facet Rasch Measurement. MESA Press.

[10] Eckes T, 2019, Many-facet Rasch Measurement: Implications for Rater-mediated Language Assessment. In Quantitative Data Analysis for Language Assessment Volume II. Routledge, London, 122–145.

[11] Han C, 2015, Investigating Rater Severity/Leniency in Interpreter Performance Testing: A Multifaceted Rasch Measurement Approach. Interpreting, 17(2): 225–251.

[12] Tseng WT, Su ZY, Nix JML, 2019, Validating Translation Test Items via the Many-facet Rasch Model. Psychological Reports, 122(2): 748–772.

[13] House J, 2015, Translation Quality Assessment: Past and Present. Routledge, London.