Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

HOW TO EVALUATE LITERARY TRANSLATIONS IN THE CLASSROOM CONTEXT: THROUGH ERROR ANALYSIS OR SCALE-BASED METHOD?

Lyu Wang

Changsha University of Science and Technology

Xiangling Wang

Hunan University

Abstract

This study evaluated students' translation quality through two different approaches to examine the effects of marking method on the assessment of literary translations in a pedagogical context. Two prose literary texts were translated by thirty-six MTI students, and then scored by nine experienced raters with error analysis and scale-based method, respectively. Scores of translations were analysed using G-studies, as conducted by the computer program *GENOVA* to compare the rater severity and consistency across marking methods. The results showed that both error-based and scale-based methods were reliable

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

tools in assessing English-Chinese literary translations, though there was a bit more rater consistency using the latter. However, a wider variation between error-based and scale-based methods was found in the assessment of Chinese-English literary translations. The variance due to raters using the scale-based method was nearly twice as much as the rater variance using the error-based method, showing there was more rater consistency using the error-based method respecting scoring leniency of rating the Chinese-English translations. Furthermore, according to the interviews, cultural and aesthetic features were highly recommended to be added as parameters in both methods for the literary translation assessment. It was also suggested that an overall consideration and rational estimation of both micro-textual and macro-textual features of literary translations might contribute to more reliable scores, regardless of rating methods and translation directions.

Keywords: the error-based method, the scale-based method, translation assessment, G-theory, English-Chinese

1. INTRODUCTION

Research with translation from native language into target language and/or the other direction has shown that translation quality assessment is both complex and problematic (House, 1997; Martínez and Hurtado Albir, 2001; Williams, 2004; Mateo, 2014; Hurtado Albir, 2015). Evidently, literary translation assessment appears to be more subjective, since literary translations entails creative interpretation based on a translator's conceptual, cultural and

aesthetic presuppositions.

Researches have shown that the assessment methods in the translation classroom can be broadly grouped into two categories: the error-based and the scale-based methods (Waddington, 2001a; Turner et al., 2010; Saldanha and O'Brien, 2014; Hurtado Albir, 2015). First, the error-based method is based on a typology of errors with a point-deduction scheme, allowing students to see how they have performed on each sentence and provides diagnostic information (Kussmaul, 1995; Lee and Ronowicz, 2014). For instance, there are 22 categories of errors in the Error Point Decisions and Framework for Standardized Error Marking of American Translators Association (ATA, 2017). Errors are marked by using a five-level fault point system (1, 2, 4, 8 and 16 points), with a total score of 180. Such a method is not as objective as they would appear. As Lai (2011) indicated, one rater may perceive an error as serious, while another may feel that it is minor or even trivial. Waddington (2001a) also stated that the choice between what is appropriate and what is inappropriate depends at least in part on the rater's personal judgement. Second, the scale-based method evaluates translation quality on the basis of rater's overall impressions, with a grading

scale (poor/insufficient/sufficient/good/excellent; from 0 to 5 or 10; etc.) (Garant, 2009; Mariana, et al. 2015; Hurtado Albir, 2015). This method treats the whole text as a unit of analysis, taking account of accuracy and fluency. It might, however, be subjective due to the varying of standards from one rater to another (Colina, 2009; Xiao, 2011; Chen, 2016).

For decades, error-based and scale-based methods have been used in translation assessment practices extensively (House, 1997; Hurtado Albir, 2015). However, determining which one to use in translation practice has been a very subjective exercise. According to a survey of translation teachers in Chinese universities (Mu, 2006), nearly fifty-one percent of eighty-five teachers surveyed used the error-based method to assess translations, while approximately fifteen percent preferred the scale-based method and thirty-four percent used both. When asked about why they chose one method over another, most of them replied that the decision was based on personal preference rather than scientific evidence from empirical research (Mu, 2006). As far as many scholars are concerned, inherent subjectivity and unreliability may be limitations to assessment procedures and

instruments. To fill the gap in empirical research into translation teaching, special attention must be paid to reliability and validity of these two methods. Hence, this article attempts to examine the reliability of rating methods in the assessment of literary translation. With generalizability (G-) theory (Cronbach, et al. 1972; Kim, 2009; Han and Huang, 2017) as the theoretical framework, it accounts for the main factors that contribute to the rating variances and compares the rater severity and consistency across marking methods. Qualitative interviews are also conducted to investigate the factors impacting raters' decision-making processes, aiming to improve the marking scales as well as the rater training on the translation assessment. Two research questions are developed in this study: 1) How different marking methods impact on literary translation assessment for direct and inverse translations? 2) What factors affect raters' decision-making in rating literary translations across marking methods? What do raters think of each method?

2. RESEARCHES ON METHODS FOR MEASURING TRANSLATION QUALITY

There has been much debate over error-based and

scale-based methods in assessing translation quality. Kussmaul (1995) pointed out that error-based assessment allowed students to see how they had performed on each sentence and provided diagnostic information. In contradiction with Kussmaul's viewpoint, Orlando (2011) insisted that translations should be evaluated holistically taking account of readability, fidelity and functions, since translation was not a simple exercise in terminology, grammar and syntactic testing. O'Brien (2012) also argued that the error-based method merely assessed quality on a segment by segment basis, giving no consideration to the larger concept of the 'context' or 'text'. Nevertheless, though scale-based method was a valid and time-saving instrument for assessing translations, teachers and students did not gain any insight into the nature of the errors encountered, "which means that it is not suitable as a diagnostic tool, or for formative purposes" (Saldanha and O'Brien, 2014: 102).

Various genres have received different level of attention in research of translation quality assessment. Although there is plenty of research into the reliability of rating methods for non-literary texts, far less attention is given to the evaluation of trainee translators' literary texts. One explanation

for this could be the booming of translation service of technical, or non-literary texts. For example, Waddington (2001b) investigated differences of the error-based method and scale-based method in assessing Spanish-English translations of an editorial from a Spanish newspaper. The result showed that the error-based method achieved a higher coefficient than the scale-based method. Unlike Waddington, Phelan (2017) pointed out the limitations of error marking scheme in assessing legal translation from English to Spanish. Her results indicated that some error categories overlapped or were vague while the flowchart for errors point decisions was difficult to implement, in particular when deciding the level of seriousness of errors. Lai (2011) compared an error-based method with three scale-based methods in rating non-literary texts, focusing on English-Chinese language pair. The results showed that the inter-rater correlations of 6/4 scales were the highest. It should be noted that the linguistic-stylistic-aesthetic features such as the extra-linguistic and contextual factors made the literary translation assessment a more complicated practice (Jureczek, 2017). Whether and how different marking methods impact on the literary translation assessment in the classroom context need to be examined empirically.

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

3. THE EMPIRICAL STUDY

3.1 The source texts

Materials in the study included source texts A (an English text) and source texts B (a Chinese text). Both of them were prose essays with a general readership, which required no specialised knowledge for the purposes of translation. In order to verify the comparability of the two texts, a strategy of multiple steps of the text selection method was adopted on the basis of Chang's (2009) and Feng (2017)'s studies. The texts were measured by the following five variables: (1) a grade level for reading, (2) total number of characters and words, (3) comprehensibility, (4) readability and (5) translatability. Firstly, six texts were selected from reading materials for students at the high school level and they were of similar length in terms of total number of characters and words. Since there is no Chinese legibility formula to measure text difficulty (Wang, 2008), comprehensibility, readability and translatability of the six texts were measured by ten professional translators based on a 0-10 Likert scale. Finally, two texts whose mean scores of text difficulty were very similar were selected.

3.2 Raters

Since our focus was the evaluation of trainee translators' literary texts, nine raters were randomly selected from nine different universities in China. They were professors, lectures and research assistants with various translation teaching background and having at least five years of experience in literary translation teaching and assessment. They varied in terms of their gender, age and rating experience, but all were highly proficient in English-Chinese and Chinese-English translation. Additionally, all of them had received formal training in assessment and translation rating.

3.3 The rating schemes

This paper focuses on two commonly-used marking schemes. Method A is based on error analysis and designed to take into account the negative effect of errors on the overall quality of the translations (Hurtado Albir, 2015; Canadian Translators, Terminologists and Interpreters Council, 2011). Errors are considered to fall into two categories in accordance with the marking scheme used in the study: translation errors and language errors (see

Table 1). Translation errors are comprehension failure to render the meaning of the original text while Language errors are expression violation of grammatical and other rules of usage in the target language. Within each category, a distinction is made between major and minor errors where the former deducts ten points from 100, whilst the latter deducts five points or three points.

Method B is a holistic method of assessment. The scale is unitary and treats the translation competence as a whole. Both English-Chinese and Chinese-English translations are rated on a five-level grading scale between 0 and 10 points, but raters consider two different aspects of students' translations: accuracy and fluency, as shown in the table 2. The grading system is as follows: 9-10 points (excellent translation), 7-8 points (good translation), 5-6 (adequate translation), 3-4 points (inadequate translation), 0-2 points (poor translation). It should be noted that for each level there are whole score (e.g. 7, 8) or half (e.g. 7.5, 8.5), in order to provide a more accurate assessment of translation skills.

Table 1: The error-based method

	Sub-categories	Examples	Value

Translation (Comprehension)	Major errors	serious misinterpretation denoting; a definite lack of comprehension of the source language; nonsense, omission of a phrase or more	-10
	Minor errors	mis-translation of a single word; omission/addition affecting meaning; lack of precision, wrong shade of meaning	-5
Language (Expression)	Major errors	gibberish, unacceptable structure	-10
	Minor errors	syntax, grammar, ambiguity, unnecessary repetition; convoluted structure, non-idiomatic structure; unacceptable translation	-5
	Minor errors	breach of spelling, punctuation; typographical conventions	-3

Table 2: The scale-based method

Band	Score	Description
5	9-10	The target text is faithful to the source text. There may be one or two lexical, grammatical, spelling or punctuation inaccuracies. The entire target text is elegant.
4	7-8	Most parts of the target text are faithful to the source text. There may be several lexical, grammatical, spelling or punctuation inaccuracies. The entire target text is readable.
3	5-6	The general idea of source text is transferred with a number of lexical, grammatical, spelling or punctuation inaccuracies. Certain parts of the target text are unreadable.

2	3-4	The original meaning of source text is partially transferred with a considerable number of lexical, grammatical, spelling or punctuation inaccuracies. Most parts of the target text are unreadable.
1	0-2	The original meaning of source text is totally inadequate transferred with continual lexical, grammatical, spelling or punctuation inaccuracies. The entire target text is completely unreadable.

3.4 The procedures

Initially, two source texts were translated by thirty-six MTI (Master of Translation and Interpreting) students as the final exam in their course of literary translation at a Chinese university. Next, nine raters from different universities participated in the evaluation workshop. A half-day rater training was conducted to reduce rater inconsistency. It was implemented at two stages, with a two-month time interval. At the first stage, rating criteria of the error-based method were clarified and discussed extensively within the rating group. A small-scale pilot study was subsequently conducted to achieve a shared understanding of the error-based system and improve the consistency of assessment. The selected nine English-Chinese translations were then scored based on the error-based method by each rater independently. Next, the other nine Chinese-English translations were assessed by the same raters with

the same method. The second stage started two months after the error-based assessment. A two-month interval was chosen to guarantee sufficient time for raters to forget the scores at the first stage, and different marking schemes also helped reduce memory interference. Another rater training for the same nine raters was conducted to ensure a thorough conceptual understanding of the scale-based method. Next, a pilot study was conducted in which raters assessed six translations with the scale-based method and discussed the rationale for scoring to reduce score variations. Subsequently, the thirty-six translations were assessed by each rater independently with the scales-based method. Afterwards, the nine raters were interviewed individually in the following two weeks to investigate their decision-making and comments on the rating methods in direct and inverse translation.

3.5 Data Analyses

To maintain consistency of score formats of the scale-based method, grades based on error analysis were converted from a 100-point scale into a 10-point scale by ten percentage. Then, G-studies were conducted using the computer program *GENOVA*

(Crick and Brennan, 1983). Within G-theory framework, translation-by-method-by-rater ($t \times m \times r$) random effects were firstly conducted. Translation-by-rater ($t \times r$) random effects were then implemented to obtain information for comparison between the error-based and the scale-based methods in assessing English-Chinese and Chinese-English translations, respectively. Finally, G-coefficients of each method were calculated to examine whether the reliability of the error-based method differed from the scale-based method for assessing English-Chinese and Chinese-English translations. In addition, raters' interviews were transcribed and analysed qualitatively.

4. QUANTITATIVE ANALYSES OF TRANSLATION QUALITY ASSESSMENT

4.1 Sources of variance in measuring direct and inverse translations

Two fully crossed design translation-by-method-by-rater ($t \times m \times r$) were conducted to explore the sources of variance in measuring direct and inverse translations. Table 3 illustrates that the sources of variance in assessing English-Chinese and Chinese-English translations presented totally different

patterns.

Table 3: Variance components analysis of G-study of $t \times m \times r$

Source of variability	Df	English-Chinese translations		Chinese-English translations	
		σ^2	% σ^2	σ^2	% σ^2
t	35	0.9022	61.38	0.0000	0.00
m	1	0.0725	4.93	0.2844	26.99
r	8	0.0332	2.26	0.0560	5.31
tm	35	0.0000	0.00	0.0987	9.37
tr	280	0.0878	5.97	0.0872	8.28
mr	8	0.0589	4.01	0.0547	5.19
tmr	280	0.3152	21.45	0.4727	44.86
total	647	1.4698	100	1.0537	100

Note. σ^2 = variance, % σ^2 = proportion of variance, *t* = translation, *m* = method, *r* = rater, *tm* = translation-by-method, *tr* = translation-by-rater, *mr* = method-by-rater, *tmr* = translation-by-method-by-rater.

As for the English-Chinese translations, the main source of variance was the translations (*t*), accounting for 61.38% in the total variance. This signified that the nine translations were substantially different in terms of quality due to differences among students' translation competence as measured by the translations. This further suggested that, as intended, the English-Chinese translations did distinguish among students if either method was used. The second largest component contributing to the total amount of variance in rating English-

Chinese translations was the residual caused by the interaction between raters, methods, translations, and other unknown covariates. Next, translation-by-rater (*tr*) was the third largest variance (5.97%), suggesting that raters did not vary substantially in assessing the English-Chinese translations. There was a smaller score variability due to the method in rating the English-Chinese translations, which accounted for 4.93% of the total variance. It could be seen that translation-by-method (*tm*) did not explain any total score variance in the group of English-Chinese translations, suggesting that there was consistency in terms of rating severity or leniency across scoring methods for assessing English-Chinese translations.

In contrast with direct translation, the greatest source of score variation in rating Chinese-English translations was the residual, accounting for 44.86% of the total variance. This indicated that the interaction between raters, methods, translations and some other unexplained sources of error posed great risk on scores, and Chinese-English translations did not accurately distinguish competent students because of the large unexplained variability. As Jiang and Wen (2010) stated, the reliability of translation test scores might also be affected by

raters of different rating experience, types of genres, and several other factors. Those hidden facets that might have contributed to the score variance should be further explored. The second largest source of variance was method (*m*), accounting for 26.99% of the total variance. This signified that there was a great difference in the scores that could be attributed to the rating method. Furthermore, the translation-by-method (*tm*) interaction component yielded the third largest variance component (9.37%), indicating that these Chinese-English translations differed a lot concerning scores across scoring methods. The value of this component (*tr*) was a little higher (8.28% of the total variance), suggesting that raters assessed these Chinese-English translations a little differently.

4.2 The impact of marking methods on translation assessment

In this section, four fully crossed design translation-by-rater ($t \times r$) were performed for the error-based and scale-based method, respectively. The purpose of these G-studies was to examine the impact of marking methods on rating direct and inverse translations. The results for both methods are presented in Table 4 and they would be further used

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

to calculate the G-coefficients in the next section.

Table 4: Variance components analysis of G-study of $t \times r$

Methods	Source of variability	Df	English-Chinese translations		Chinese-English translations	
			σ^2	% σ^2	σ^2	% σ^2
Error-based method	*t*	35	0.8460	65.69	0.1354	19.44
	r	8	0.1146	8.90	0.0749	10.76
	tr	280	0.3273	25.41	0.4861	69.80
	total	323	1.2879	100	0.6964	100
Scale-based method	*t*	35	0.9415	63.20	0.0206	2.57
	r	8	0.0695	4.67	0.1464	18.29
	tr	280	0.4787	32.13	0.6337	79.14
	total	323	1.4897	100	0.8007	100

Note. σ^2 = variance, % σ^2 = proportion of variance, *t* = translation, *r* = rater, *tr* = translation-by-rater.

In the group of English-Chinese translations, the discrepancy between error-based and scale-based assessment is relatively small, with the same ranking of the variance components yielded by three objects of measurement (*t*, *tr*, *r*). At first, for both error-based and scale-based methods, the main sources of variance were the translations (*t*), reflecting individual translation competence accounted for more than half the variance. In this case, the purpose of evaluation, to detect differences between examinees, was achieved. The next largest variance components of both methods were the residual,

which contained the variability due to the interaction between raters and translations, and other unexplained sources of variance. The last variance component was the rater (*r*), with a nearly twice larger proportion for the error-based method than that for the scale-based method. It revealed that there was more rater inconsistency using the error-based method concerning scoring leniency of assessing English-Chinese translations. This result has been confirmed by several research studies (Xiao, 2012; Waddington, 2001a), although it is contradictory to the findings as reported by Lai (2011), who claimed that in both English-Chinese and Chinese-English translations, the inter-rater correlations of 6/4 scales were higher, compared with error analysis.

Regarding Chinese-English translation, there was a wide variation between error-based and scale-based methods. Initially, the largest variance components for both the error-based and scale-based method were the residual (*tr*), accounting for 69.80% and 79.14%, respectively. These large proportions indicated that the scores of Chinese-English translations differed substantially across raters and there were more variabilities using scale-based assessment, compared with the error-based method.

Moreover, differences between error-based and scale-based assessments are substantially greater when looking at the variance components yielded by translation (*t*) at 19.44% and 2.57%, respectively. Finally, the variance explained by the rater (*r*) (18.29% of the total variance) using the scale-based method was nearly double in proportion to that of error-based method. It revealed that there was more rater inconsistency using the scale-based method than that using the error-based method respecting scoring leniency of rating the Chinese-English translations.

In summary, both error-based and scale-based methods were reliable tools in assessing English-Chinese translations, though there was a bit more rater consistency using scale-based method. However, a wider variation between error-based and scale-based methods was found in the Chinese-English translation assessment. The variance due to raters (*r*) using the scale-based method was nearly twice as much as the rater variance using the error-based method, showing there was more rater consistency using the error-based method respecting scoring leniency of rating the Chinese-English translations. It also indicated that the error-based method appeared to be more sensitive to differences

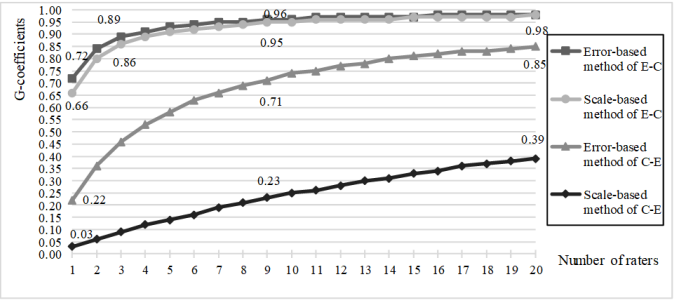
among Chinese-English translations than the scale-based method did. According to the raters' interviews, the error-based method generated more information for each sentence than the scale-based method did, which might lead to greater precision of measurement in assessing Chinese-English translations. Raters also acknowledged that identical scores would be assigned to Chinese-English translations more often by using the scale-based method.

4.3 Measuring the reliability of each method

In this part, the G-coefficients were calculated for each method to estimate the reliability in both translation directions according to the varying numbers of raters (see Figure 1). As demonstrated in Figure 1, with the number of raters increasing from one to nine, the G-coefficients of each method improved at an increasing rate. From nine to twenty raters, the generalizability coefficient continued to increase, but at a constant rate. However, the G-coefficients of each method showed a big discrepancy in different translation directions. The G-coefficients of the scale-based method followed the same trend as the error-based method in assessing English-Chinese translations, while the G-

coefficients of the two methods showed considerable differences in the evaluation of Chinese-English translations.

Figure 1. Summary of dependability G-coefficients



Note. *E-C* stands for English-Chinese translation; *C-E* stands for Chinese-English translation.

The G-coefficient of English-Chinese translations obtained for the error-based was a little higher than that for the scale-based method in the current nine-rater scenario (.96 and .95 respectively). Besides, the G-coefficients for both methods with three raters were above .85, indicating that it was sufficiently reliable to assessing English-Chinese translations. This result corresponded to Xiao's (2012) finding that three raters achieved very high inter-rater reliability in assess English-Chinese literary translations. If the number of raters was increased to 20 in the group of English-Chinese translations, the

G-coefficient for both the error-based and the scale-based method would be .98. Accordingly, scale-based method could yield as reliable and dependable results as the error-based method for marking English-Chinese translations.

Nevertheless, in the group of Chinese-English translations, when nine raters were involved in the design, the G-coefficient for the error-based method was .71, whereas the G-coefficient for the scale-based method was only .23. If the number of raters was increased to 20, the G-coefficient for the error-based would be .85, however, the G-coefficient for the scale-based method was merely .39, much lower than that of .71 obtained by the error-based method with nine raters. These results indicated that there was a rather low reliability using the scale-based method in assessing Chinese-English translations. Furthermore, they would raise concerns about the reliability of translation assessment in the classroom assessment context, where only one rater evaluates Chinese-English translation (Xiao, 2011). As shown in Figure One, for a design with one rater, the G-coefficient for error-based method was .22, while that for scale-based method was only .03. Compared with English-Chinese translation, the reliability of both methods in assessing Chinese-English

translation was much lower, even if the number of raters increased to twenty. One reason for this might be that levels of translation quality would impact the scoring reliability of translation assessment. Just as Chen (2016) mentioned in her study, there was more absolute consistency in rating Chinese-English translations of high quality than those of poor quality. Since the translators in this study were first-year MTI students with Chinese as their native language, their translations from Chinese to English were unsatisfactory. The factor concerning levels of translation quality needed further investigation. Besides, this result also indicated that more reliable, effective and pragmatic measurement of Chinese-English translations should be developed if the number of raters was taken into consideration.

5. QUALITATIVE ANALYSES OF RATERS' DECISION-MAKING

In order to further examine the literary translation assessment, semi-structured interviews were conducted to investigate the factors impacting raters' decision-making when using different marking methods. It can help us better understand the reason behind raters' decision-making, and finally contribute to developing a more reliable

evaluation of literary translations. The semi-structured interview method combined three structured questions with some unstructured exploration. The three questions were as follows: 1. What makes a good literary translation? 2. In what ways each of the factors you mentioned affect your rating when using different methods in assessing direct and inverse translations? 3. How do you think of each method? Do you have any suggestions?

According to the interviews, we detected three factors that had an impact on raters' decision-making in literary translation assessment, namely the linguistic factor, the cultural factor and the aesthetic factor. First, raters explained that when rating English-Chinese translations using the scale-based method, their focus was generally on macro-textual level, i.e. on the aesthetic level of the form and content. The raters agreed that their decision-making regarding rating was greatly influenced by the artistic conception, lingering charm of the text, aesthetic empathy and emotion in the Chinese text. They also highlighted the impact of formal aesthetic markers such as choice of words and rhetorical devices on their decision-making. However, they might not take into account the minor linguistic errors because they held that the macro-textual

equivalence can undermine some micro-textual errors. As an illustration, it is worth highlighting a comment made by Rater A with respect to his decision-making process as follows:

> When rating English-Chinese translations with the scale-based method, I think it is very important to present the artistic conception, lingering charm and aesthetic empathy of the original text. Thus, I focused on the aesthetic factors of the translation instead of the mistranslations in the linguistic forms. Because mistranslations which appeared in the compensation of the function and effect of the source text might help to obtain a functional equivalence. (Rater A; Our translation)

This is consistent with Rodríguez's (2007) proposal that slight divergences are accepted in accounting for the functional equivalence of the text. Similarly, Läuscher (2000) proposes the notion of flexibility, applying it specifically to literary translation evaluation.

In the process of rating English-Chinese translations with the error-based method, considerable attention had been paid to both the macro-textual and micro-textual features of translations. All raters agreed that

they focused on the judging and recording of linguistic errors, due to the nature of the error analysis rubric. Those linguistic factors, including information integrity, information accuracy, readability and style/register impacted their decision-making greatly. Furthermore, they would also consider the macro-textual features of translations, which gave them a general impression of the quality of translated texts:

> Although rating English-Chinese within the framework of error-based method, I will estimate the overall quality of the translation based on the linguistic features, choice of words and context before detecting translation errors and deducting points. If the result is far different from the estimated score, I will re-evaluate the translation. (Rater B; Our translation)

This was in line with the rating procedures proposed by Waddington (2001b).

In the different translation direction, raters reported that linguistic factors and formal aesthetic factors had the greatest impact on their decision-making using either method. Firstly, when rating translations with the scale-based method, raters' decision making was largely motivated by their

consideration for three factors: syntactical structures, the choice of words and grammatical correctness. It was believed that slightly greater weight was placed on syntactical structures than that on the other two, even though the effect of three-way interactions cannot be ruled out. This phenomenon was also discovered in Chen's (2016) study. Raters presented a similar performance in rating Chinese-English translation using the error-based method, putting more effort on the spelling and punctuation. Secondly, raters had inconsistent perceptions toward linguistic and cultural features of translations using different rating methods. In fact, raters acknowledged that they tended to pay closer attention to cultural factors when using the error-based method to assess Chinese-English translations. One rater explained that:

> If the same meaning of a culture-specific item was not successfully transferred into the target text, I would mark it as a minor linguistic error. However, I might not perceive it as an error when assessing the translation holistically. (Rater B; Our translation)

This was also found by Lee and Ronowicz (2014). Additionally, the raters also indicated that it was important for the translator to detect the implied Chinese meaning and bring it out explicitly in the English version, because English is an overt

cohesion-prominent, hypotactic language and emphasizes language form to achieve cohesion while Chinese relies more on semantic linkage and implicit logic for the same purpose.

> When translating from Chinese to English, it is necessary to explicitly reproduce the implied logic relations within and between sentences in the target text. The loss of implied meaning will be marked as a major error and deducted five point. (Rater C; Our translation)

In addition, raters were asked to make their comments on the error-based and the scale-based methods, and provide a few suggestions for modifications of the two rating systems. Most raters acknowledged that they would like to adopt the scale-based method to rate English-Chinese translations. They found it more convenient and appropriate to treat the text as the unit of assessment and rated it on the basis of their global impression or overall judgment. They also complained that the error-based method had too many parameters, which broke down the wholeness of the text and made the evaluation process time-consuming. For example, one of the raters stated that:

> It's challenging if I have to consider time. Scoring becomes very difficult when there

> are a large number of translations to be assessed sentence by sentence in a limited time. Besides, error-based standards are too complex, so the assessment procedures become tiring. (Rater E; Our translation)

Although the statistical analysis indicates the superiority of error-based method over the scale-based method, raters revealed the limitations of the error-based method, and the benefits of combining both methods. As an illustration, one rater claimed that:

> I think both the English-Chinese and Chinese-English translation should not be assessed simply as right or wrong. Translation, literary translation in particular, is a rather creative activity, which is also distinguished by its aesthetics and communicative function. Combining the two methods may compensate for limitations in each method. (Rater I; Our translation)

Accordingly, when rating translations based on error analysis, raters also took the macro-textual features into account. This phenomenon was also observed by Mu (2006). Given the approaches to making the rating process more reliable, the raters agreed that the combination of both methods would present greater accuracy. The research by

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

Waddington (2001b) provided evidence in favour of the benefits of combining error analysis and holistic methods in rating Spanish-English translations. With regard to the suggestions for the improvement of the two methods, raters highly recommended that cultural and aesthetic features should be added as parameters in rating literary translations. They also proposed that the assessment of literary translation should be grounded on the consideration of the text-types, functions, coherence and cohesion as well as the purpose, acceptability, and intertextuality of target texts. The results might hold the potential to improve the assessing system as well as the rater training on literary translation assessment.

6. CONCLUSIONS

This study was initiated to investigate score reliability for a rater-mediated assessment of literary translations, using generalizability theory and qualitative data from the interviews. As for English-Chinese translations assessment, the results indicated that there was a smaller variance due to the assessment method. The G-coefficients that were obtained for the error-based method were very similar to those obtained for the scale-based method. When it came to the Chinese-English translations, there was more rater inconsistency using the scale-

based method than that using the error-based method respecting scoring leniency. Besides, the result of G-coefficients showed that the reliability of the scale-based method was relatively lower than that of the error-based method for marking Chinese-English translations. Furthermore, according to the interviews, cultural and aesthetic features were highly recommended to be added as parameters in both methods for the literary translation assessment. An overall consideration of both micro-textual and macro-textual features of translated texts might contribute to more reliable scores.

There are three main limitations to this study. To start with, given that the error-based method and the scale-based method are merely two of the scoring criteria available, other scoring criteria might lead to different results. Next, this study only analyses the reliability of literary translation assessment, however, other text types such as technical translation might affect the rating reliability. Then, some other hidden factors such as levels of translation quality, text difficulty and raters' L2 background are not taken into considerations in this study, which might contribute to score variance. Despite these limitations, the current study suggests several implications for translation teaching and

future research. On the one hand, the results show that the scale-based method could be used reliably as the error-based method for assessing English-Chinese literary translations in the classroom context, if raters are well trained in using the scales. On the other hand, the lower reliability of the scale-based method in marking Chinese-English literary translations remains alarming. Closer examination is needed to explore whether a combined method could increase the reliability and the best way to combined these two methods in the assessment of Chinese-English literary translations.

ACKNOWLEDGEMENT

Thanks go to our participants and raters for their precious time. Particular gratitude is extended to Professor Milo Kaufmann and Helen Kaufmann for their comments on the earlier drafts of this article. This research was supported by the National Social Science Fund of China (ref. 2019BYY104).

REFERENCES

American Translators Association. (2017). *Explanation of Error Categories*. https://www.atanet.org/certification/aboutexams_error.php

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_2020_9_lyu_wang_and_xiangling_wang.pdf*

Chang, C.Y. (2009). *Testing applicability of eye-tracking and fMRI to translation and interpreting studies: an investigation into directionality* [Doctoral dissertation, Imperial College London]. EThOS. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.508786

Chen, Y. (2016). Yingyu zhuanye baji kaoshi hanyiying xiangmu pingfenyuan yanjiu—yi 2015 nian TEM8 weili [An examination of rater performance on the Chinese-to-English translation section of TEM8 in 2015]. *Waiyu dianhua jiaoxue*, *171*(5), 77-82.

Colina, S. (2009). Further Evidence for a functionalist approach to translation quality evaluation. *Target*, *21*(2), 235-264. https://doi.org/10.1075/target.21.2.02col

Crick, J. E. & Brennan, R. L. (1983). *Manual for GENOVA: a generalized analysis of variance system* (American College Testing Technical Bulletin NO. 43). Iowa City, IA: American College Testing, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley.

Canadian Translators, Terminologists and Interpreters Council. (2011, February). *Standard certification translation examination marker's guide*. http://www.cttic.org/examDocs/guide.markersE.pdf

Wang, L. and Wang, X. (2020). How to evaluate literary translations in the classroom context: through error analysis or scale-based method? *Current Trends in Translation Teaching and Learning E, 7, 276-313. 10.51287/cttl_e_20 20_9_lyu_wang_and_xiangling_wang.pdf*

Feng, J. (2017). Yiru yichu renzhi fuhe bijiao yanjiu [Comparing cognitive load in L1 and L2 translation: evidence from eye-tracking]. *Zhongguo waiyu*, *14*(4), 79-91.

Garant, M. (2009). A case for holistic translation assessment. *AFinLA-e Soveltavan kielitieteen tutkimuksia*, *1*, 5-17.

Han, T. & Huang, J. Y. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: a Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, *53*, 112-147.

House. J. (1997). *Translation quality assessment: a model revisited*. Gunter Narr.

Hurtado, Albir. A. (2015). The acquisition of translation competence. competences, tasks, and assessment in translator training. *Meta*, *60*(2), 256-280. https://id.erudit.org/iderudit/1032857ar

Jiang, J. L. & Wen, Q. F. (2010). Jiyu Rasch moxing de fanyi ceshi xiaodu yanjiu [Validation of a translation exam based on Rash Measurement Model]. *Waiyu dianhua jiaoxue*, *131*, 14-18.

Jureczek, P. (2017). Literary translation quality assessment: an approach based on Roland Barthes' five literary codes. *TranslatoLogica: A Journal of Translation, Language, and Literature*, *1*, 136-155.

Kim, Y. (2009). A G-Theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics*,

30(3), 435–440. https://doi.org/10.1093/applin/am p035

Kussmaul, P. (1995). *Training the translator*. John Benjamins.

Mariana, V., Cox, T. & Melby, A. (2015). The multidimensional quality metrics (MQM) framework: a new framework for translation quality assessment. *The Journal of Specialised Translation*, *23*, 137-161. https://www.jostrans. org/issue23/art_melby.pdf

Martínez, M. N. & Hurtado Albir, A. (2001). Assessment in translation studies: research needs. *Meta*, *46* (2), 272–287. https://id.erudit.org/iderudit/003624ar

Mateo, R. M. (2014). A deeper look into metrics for translation quality assessment (TQA): a case study. *Miscelánea: A Journal of English and American Studies*, *49*, 73–94. https://www.misce laneajournal.net/index.php/misc/article/view/170

Mu, L. (2006). Fanyi ceshi ji qi pingfen wenti [Translation testing and grading]. *Waiyu jiaoxuei yu yanjiu*, *38*(6), 466-480.

O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, *17*, 55-77. https://www.jostrans.org/ issue17/art_obrien.pdf

Orlando, M. (2011). Evaluation of translations in the training of professional translators: at the crossroads between theoretical, professional and pedagogical practices. *The Interpreter and*

Translator Trainer, *5*(2), 293-308. https://doi.org/ 10.1080/13556509.2011.10798822

Phelan, M. (2017). Analytical assessment of legal translation: a case study using the American Translators Association framework. *The Journal of Specialised Translation*, *27*, 189-210. http:// www.jostrans.org/issue27/art_phelan.pdf

Rodríguez, B. M. (2007). *Literary translation quality assessment*. Lincom Europa.

Saldanha, G. & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.

Turner, B., Lai M. & Huang, N. (2010). Error deduction and descriptors: A comparison of two methods of translation test assessment. *Translation & Interpreting*, *2*(1), 11–23. http://www.transint. org/index.php/transint/article/view/42

Lai, T. Y. (2011): Reliability and validity of a scale-based assessment for translation tests. *Meta*, *56*(3), 713–722. https://id.erudit.org/iderudit/1008341ar

Läuscher, S. (2000). Translation quality assessment: Where can theory and practice meet? *The translator*, *6*(2), 149-168. https://doi.org/10.1080/ 13556509.2000.10799063

Lee, Y. O. & Ronowicz, E. (2014). The development of an error typology to assess translation from English into Korean in class. *Babel*, *60*(1), 35-51. https://doi.org/10.1075/babel.60.1.03lee

Waddington, C. (2001a). Different methods of evaluating student translations: question of validity. *Meta*,

46(2), 311-325. https://id.erudit.org/iderudit/004583ar

Waddington, C. (2001b). Should translations be assessed holistically or through error analysis? *Hermes, Journal of Linguistics*, *26*, 15-37. https://doi.org/10.7146/hjlcb.v14i26.25637

Wang, L. (2008). Keduxing gongshi de neihan ji yanjiu fanshi [Some concepts of readability formula and relevant research paradigm as well as the research tasks of formula in TCFL]. *Yuyan jiaoxue yu yanjiu*, *6*, 46-53.

Williams, M. (2004). *Translation Quality Assessment: An Argumentation-Centred Approach*. University of Ottawa Press.

Williams, M. (2009). Translation Quality Assessment. *Mutatis Mutandis*, *2*(1), 3-23.

Williams, M. (2013). A holistic-componential model for assessing translation student performance and competency. *Mutatis Mutandis, 6*(2), 419-443.

Xiao, W. Q. (2011). Fanyi ceshi de pingfenyuan xindu yanjiu [Rater Reliability of Translation Testing]. *Waiyu xuekan*, *6*, 115-119.

Xiao, W. Q. (2012). Butong pingfen fangfa xia fanyi ceshi pingfenyuan jian xindu de shizheng yanjiu [On the Inter-rater Reliability of Translation Tests of Above-medium Length for College Students]. *Jiefangjun waiguoyu xueyuan xuebao*, *35*(4), 46-50.