LUBLIN STUDIES IN MODERN LANGUAGES AND LITERATURE
Maria Curie-Sklodowska University Press
Vol 48, No 3 (2024)
E-ISSN 2450-4580

Nilufer Aybirdi, Bayburt University, Turkey
Turgay Han, Ordu University, Turkey

DOI:10.17951/lsmll.2024.48.3.89-109

An Investigation of Rater Effects on L2 Translation Performance Scores

ABSTRACT
This study examines the impact of L2 translation quality and rating experience on raters' scoring behaviours and the reliability and variability of scores. It also investigates frequently used decision-making strategies applied by raters assigning scores to different qualities of translation papers produced by Turkish EFL students. In total, 80 translation papers (40 low-quality and 40 high-quality) obtained from the participants were given to 10 raters to score using a translation scoring rubric. Results revealed that less experienced raters were more positive while scoring the students' translation performances and assigned higher and more consistent scores to the papers.

KEYWORDS
translation assessment; decision-making strategies; rating experience; translation quality; think-aloud protocols

1. Introduction

The use of translation as a writing assessment tool in language testing has gained increased popularity and has become an applicable method in teaching foreign languages to students (Cook, 2010). Translation can be evaluated as a writing performance since scorers grade a translation text according to features such as linguistics, content, style, organization, and various technical aspects (Marais, 2013). Assessing second language (L2) learners' translation performance, however, can also be a problematic and rigorous task when, for instance, the effect of the different social contexts on writing processes (Baker, 2010), and the impact of L2 learners' background knowledge, language proficiency level and judgmental ability on the writing performance (Heaton, 2003), are taken into account.

Nilufer Aybirdi, Department of Foreign Languages, Department of English Language Teaching, Faculty of Education, Bayburt University, Bâbertî Külliyesi Tuzcuzade Mahallesi Demirözü Caddesi Erzincan Yolu Üzeri 2.7 km, Bayburt, Phone: 00905413705535, niluferaybirdi@hotmail.com, https://orcid.org/0000-0002-1223-3050
Turgay Han, Department of English Language and Literature, Faculty of Sciences and Literature, Ordu University, Cumhuriyet Yerleskesi, Cumhuriyet Mahallesi, 52200, Altınordu, Phone: 00905051248512, turgayhan@yahoo.com.tr, https://orcid.org/0000-0002-9196-0618

The journal is financed by the Institute of Modern Languages and Literatures of Maria Curie-Sklodowska University.

When raters judge the quality of translated texts by students, several factors can affect the scoring procedure found in the assessment of L2 writing performance. The first of these factors is related to the characteristics of the raters. The background language knowledge (Chang, 2002; Pöchhacker, 1994), decision-making strategies (Baker, 2010), preferred scoring methods (Barkaoui, 2007), the tendency of severity or leniency (Huang, 2008), and the previous rating experience of the raters (Pöchhacker, 1994; Şahan, 2018), are among the factors that impact how a translated text is judged. The second factor that impacts the scoring process is the type of rubric (e.g., holistic or analytic) used by the raters (Barkaoui, 2007; Yıldız, 2020) and preferences for different rating procedures may affect the variance of scoring (Chang, 2002). The other factors that affect scoring are the quality of the translated texts produced by the students and their L1 proficiency (Şahan, 2018)

2. Literature

In the field of writing assessment, some studies have attempted to determine the impact of text quality on the reliability of raters' scores (Brown, 1991; Ferris, 1994; Han, 2017; Huang, 2008; Şahan, 2018). For example, using Generalizability Theory (G-theory), Han (2017) compared scores assigned to three different qualities of essays (low, medium, and high) by raters with different previous rating experiences. The results of the study showed that while the raters were similar in the scores they assigned to high-quality papers, they were significantly different in the scores they gave to low-quality papers. Finally, yet another factor is related to the test-takers (i.e., students). L2 students generally receive lower writing performance scores than native English students when they are asked to write essays on a topic (Huang, 2008). Students' background knowledge and target language proficiency impact their understanding and interpretation of the writing tasks (Han, 2017). Similarly, in cases where students are asked to translate from the target language to the native language, or in reverse, the competencies of the students in both languages may affect the quality of the written products. According to Baba (2009), the ability to appropriately express words, phrases, and idioms in an L2 contributes to students' writing performances and the scores assigned to them by the raters.

The presence of human (teacher) interference in the process of scoring translation tests makes it highly subjective which is indicative of unreliability (Lado, 1964). When a person carries out the scoring of a translation performance, several factors impact the process of translation performance assessment, causing a potentially subjective and inconsistent assessment. Rater subjectivity (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education [AERA, APA, and the NCME], 2014) is one of the sources of reliability, as it is well known that raters may assign

different scores to the same essay (i.e., inter-rater reliability), or that the same rater may assign different scores to essays which are of the same quality (i.e., intra-rater reliability), which may threaten the reliability of the scores (Brown, 2004; Homburg, 1984). A higher degree of reliability should be ensured when the test scores are used to make high-stakes decisions that are not easily reversed (AERA, APA, and the NCME, 2014). Thus, it is assumed that any difference between an individual's scores obtained at different times regarding the same measurement situation may have resulted from one or multiple sources of error rather than from the individual's maturation or learning (Güler et al., 2012). A reliable and objective measurement of translation quality is therefore essential in an academic setting in educational programs for formative evaluation. This may include eliminating applicants during procedures of admission, giving feedback on learner progress and performance, and testing what they have acquired at the end of a program (Angelelli, 2009).

3. Empirical studies on translation assessment

In the 1990s and early 2000s, some scholars claimed that there was a scarcity of empirical research in the field of L2 translation assessment studies (e.g., Hatim & Mason, 1997; Melis & Albir, 2001; Pym, 1992). They indicated that although a few studies had been carried out on translation assessment, they had been conducted neither objectively nor in a regimented fashion (Melis & Albir, 2001, p. 273). A theoretical and descriptive approach has been applied to the issue of translation assessment, even in the latest publications. This situation clearly exposes the absence of empirical studies in the area of translation assessment. From the 1990s to the present time, the amount of empirical research carried out in the field of L2 translation quality assessment is very limited. Some studies have investigated the use of translation as an assessment toolkit in academic settings (e.g., Calis & Dikilitas, 2012; Källkvist, 1998; Laufer & Girsai, 2008; Lee, 2013; Prince, 1996), whilst others have focused on determining the impact of translation on L2 learners' accuracy (e.g., Berggren, 1972; Ghaiyoomian & Zarei, 2015; Källkvist, 2008; Mundt & Groves, 2016; Soleimani & Heidarikia, 2017; Stapleton & Kin, 2019; Uzawa, 1996; Vaezi & Mirzaei, 2007). Studies searching for an objective method to assess translated works are even scarcer in the existing literature, primarily focusing on the examination of the reliability and validity of developed translation tests (e.g., Colina, 2008; El-Banna, 1993; Eyckmans et al., 2009; Ghonsooly, 1993; Han & Shang, 2023; Ito, 2004; Neves, 2002; Orozco, 2000; Tavakoli et al., 2012).

In the literature, to the best knowledge of the authors, no research has yet been conducted to examine the impact of rating experience on L2 translation score variability and the issue of the reliability of scores in L2 translation assessments. This study aims to fill this gap by examining issues of rater reliability in L2

translation assessment in the context of higher education. Furthermore, no research has used G-theory to determine sources of score variability, and the use of thinking-aloud protocols (TAPs) has not been widely preferred by researchers since it is a challenging task to collect, prepare (transcription) and analyse verbal data. Therefore, this study is assumed to contribute to the translation assessment literature by investigating rater cognition through verbal protocols.

4. Research questions

The following research questions have guided this study: 1) What are the sources of score variation that contribute to the score variability of the scores assigned to high- and low-quality translation papers? 2) Are there any significant differences between the scores assigned to low- and high-quality translations? 3) Does rating experience have an impact on the variability and reliability of the scores assigned to high- and low-quality translation papers? 4) How do raters make decisions while assigning scores to translation papers of different quality?

5. Method

5.1. Design of the study

Following a mixed-methods research design, the data was collected both quantitatively and qualitatively. Quantitative data was obtained from the comparison of the variability of the scores assigned by the raters to the translation texts (papers). The qualitative data was obtained from the recordings of think-aloud protocols (TAPs). Official permission was also obtained from the Dean's Office of the Faculty, where the students were enrolled.

5.2. Selection of raters

A total of ten raters (four females and six males) participated in the study voluntarily. The raters had more than one year of experience in teaching and evaluating English language and were professionals in the field of interdisciplinary English language teaching, learning, and assessment. Prior to the main data collection, an adapted rater profile form (Barkaoui, 2007; Cumming et al., 2002) was given to the raters to obtain information about their personal, educational, and professional backgrounds. Five participant raters reported five years or less experience in rating translation papers of EFL students and were categorized as less experienced raters. The other five participant raters declared they have seven years or more of grading experience and were categorized as more experienced raters.

5.3. Selection of EFL students

In selecting the students, the researchers followed a purposive sampling method. The students who participated in the study were majoring at the English Language and Literature Department at a state university in Türkiye. Among 115 sophomore

students, 40 volunteers (thirty females and ten males) were included in the study. Before their first year, all of these students had received foundational courses for one year in the department. On completion, they were administered a language proficiency level exam which they all successfully completed. For this reason, they were all assumed to have a minimum level of B2 language competency. For the treatment, necessary permission was obtained from Kafkas University (E.1900066389). A written informed consent form was also provided to the students affirming that they had the right to discontinue their participation at any time should they so wish.

5.4. Data collection instruments

The quantitative data was collected through scores given to translated texts, and the qualitative data was collected through a background questionnaire and TAPs.

5.4.1. Translation passages

In this study, four informative newspaper articles obtained from an online British magazine (The Daily Star Online) were selected to be translated by the students in four different sessions, controlled by the first author of this paper. The Flesch-Kincaid Test showed that the chosen texts had scores between 62.5 and 69.96 for Flesch Reading Ease, indicating that the selected texts were suitable for the language proficiency level of the students. In each translation session, all 40 students were asked to translate a newspaper article from English to Turkish. According to Dickins et al., (2016), training translators with a focus on translation into a native language results in a higher quality translation compared to a translation from a native language into the target language. The students were given 90 minutes to complete each translation task, which were performed by hand with pen and paper.

Overall, 160 translation papers were obtained from the students who participated in the study. The authors of this study and an expert rater carefully divided the collected translation papers in accordance with quality (e.g., high and low), following the criteria of the 10-point rating scale described below. While the papers graded over five were categorized as high quality, the ones that were graded under five were categorized as low quality. At the end of this procedure, 40 high-quality and 40 low-quality translation papers were randomly selected for analysis.

5.4.2. The 10-point rating scale

The 10-point rating scale used in this study was developed by Marais (2013) for the purpose of evaluating student translation products in educational settings. Although the developer of the rubric labelled the rubric as holistic, it presents features of an analytic rubric in that it consists of six sections structured hierarchically: 1)

suitability of translation for the general purpose, 2) culture and target reader, 3) text, 4) design, 5) content and finally, 6) language assessment. In this hierarchy, the sections are not ranked in order of importance: the principal objective is to reflect the decisions taken by the students during their translation process. Performance scores were assigned to the subcategories of rubric sections as follows: suitability of translation for the general purpose (1 pt.), culture and target reader (2 pts.), text (1.5 pts.), design (1.5 pts.), content (1.5 pts.) and language (2.5 pts). The internal consistency of the evaluation rubric was established as 0.89. In the current study, this scoring rubric was preferred for its simplicity and assumed to be used as a scoring system. Table 1 displays the score weight distribution of the six components of the rubric used in the study.

Table 1. The score weights of six categories in the 10-point rating scale

Category	Weight Percentage
Purpose	10 %
Culture	20 %
Text	15 %
Technical Aspects	15 %
Content	15 %
Language	25

5.4.3. The Think-Aloud protocols (TAPs)

By including TAPs in this study, the researchers aimed to investigate the raters' internal decision-making process while they evaluated the translation papers. Studies including TAPs specifically emphasize that affective factors significantly impact translation assessment (Laukkanen, 1996). In this study, the raters were asked to use TAPs while scoring pre-determined (randomly) 32 of the 80 translation texts that were in their translation paper pack with a voice-recording device. Raters were asked to provide an accurate, clear, and consistent report of this cognitive process through the application of the introspection method in the evaluation process. These recordings were then transcribed, and raters' evaluation-based utterances were centred upon this information. In the analysis of this data, the researchers made the coding and categorization. In order to ensure that the raters fully understood the think-aloud procedure, the researchers followed a TAP training process which included a training session for the raters, focusing on the application of this method as well as the rubric.

5.4.4. Data analysis

Descriptive and inferential analyses were conducted on both the total scores assigned to the translation papers and on the sub-scores given to six components of the papers. In addition, descriptive statistics were applied to the codes obtained

from the TAPs' analysis. The G-theory framework was used to determine independent variation sources and identify score variation due to experience. While descriptive and inferential statistical analyses were performed using SPSS, the G-study analyses were performed with the computer program EDUG. Analysis of the data obtained from the TAPs was carried out by application of a coding scheme adapted from Cumming, Kantor, and Powers (2002). The qualitative data set was composed of the TAPs the raters recorded for 32 randomly selected translation papers while evaluating them. Following a top-down approach, the TAPs were divided into meaningful units according to three criteria a) by the rater reading aloud a section of the translation paper, b) by the rater commenting about how the translation of a sentence should be and c) when the rater stated a clear idea/thought about the translation paper by adopting a holistic manner. After the segmentation of the data procedure had been completed, a discussion session was held with two field experts regarding each item included in the coding scheme. Then, by consulting another field expert, the sub-categories that the coding frame included were reevaluated and modified by the researcher. To ensure the inter-rater reliability of the coding system, another expert who had knowledge of qualitative data analysis was asked to code a randomly chosen sample of 15% of the protocols. The analysis carried out to determine the similarity between the coding procedures indicated an agreement of .82.

5.4.5. G-theory framework

In the literature, three different theoretical frameworks have been applied in the research of L2 performance assessment to examine measurement reliability: Classical Testing Theory (CTT), Generalizability Theory (G-theory), and IRT (Brennan, 2011). Among these performance assessment frameworks, G-theory (developed by Cronbach e al., 1972) arose due to the limitations of CTT, back in 1972. In comparison to CTT, which centres on estimating only a single error of measurement at a time (e.g., item, rater, form, etc.), G-theory enables researchers to analyse multiple sources of error variance simultaneously (e.g., raters, tasks, topics) (Brennan, 2001). In the examination of the reliability of behavioural measurements, G-theory ensures a very practical and flexible framework (Shavelson et al., 1989). It determines every source of systematic and unsystematic error, distinguishes them, and estimates each one of them (Webb & Shavelson, 2005).

For these reasons, G-theory was applied as a methodological framework in the current study. Within the G-theory framework, further data analyses were conducted using the following three steps:

1) A paper-by-rater-by-quality ($p \times r \times q$) random effects G-study was employed in order to determine independent sources of variation such as papers (p), raters (r), quality (q), paper-by-rater ($p \times r$), paper-by-

quality (*p x q*), rater-by-quality (*r x q*), and paper-by-rater-by-quality (*p x r x q*) for 80 translation papers evaluated by holistic scoring methods. Also, calculations of generalizability and dependability coefficients were performed to determine the reliability of the data set.

2) In order to attain variance component estimates, a paper-by-rater (*p x r*) random effects G-studies were employed separately; papers (*p*), rater (*r*), paper-by-rater (*p x r*) for 40 low-quality and 40 high-quality translation papers scored by holistic evaluation methods. Also, calculations of generalizability and dependability coefficients were performed to determine the reliability of the data set.
3) Since raters were categorized into two groups with respect to their previous rating experience as less experienced and more experienced, the scores they assigned to the translation papers were compared in terms of their generalizability and dependability coefficients. Thus, a paper-by-rater-by-quality (*p x r x q*) random effects G-study was performed on all translation papers, and a paper-by-experience (*p x r*) random effects G-study was conducted on high- and low-quality translation papers.

6. Results

6.1. Results for the first research question

What are the sources of score variation that contribute to the score variability of the scores assigned to high- and low-quality translation papers?

Table 2. Variance components for random effects P X R X Q design

Variance source	df	σ^2	%
P	39	4.84	61.8
R	9	-0.33	0.0
Q	1	-0.03	0.0
PR	351	0.98	12.6
PQ	39	0.10	1.4
RQ	9	0.94	12.1
PRQ	351	0.95	12.1
Total	799	-	100

In order to identify the sources of the variance, a paper-by-rater-by-quality *(p x r x q)* random effects G-study was carried out. Components of the variance and their relative contribution to the variability of scores are presented in Table 2.

Table 2 indicates that the largest variance component was papers (61.8%), indicating that students had considerably different translation performances. The second greatest variance that contributed to the score variability was the interaction between papers and raters (12.6%), indicating that raters differed largely from one to another in terms of scoring the students' translation. The third greatest variance

was attributable to the residual (12.1%), indicating that the variability is due to the interaction between raters, translation quality, paper, and other unexplained systematic and unsystematic sources of error. The fourth greatest variance contributing to the score variability was the interaction between the raters and translation quality (12.1%), indicating considerable consistency in raters' ratings across paper quality. The remaining sources of variance such as rater, translation quality, and paper by quality were determined to contribute relatively little to the variability of the scores (0%, 0%, and 1.4% respectively).

To identify the sources of variance that contribute to the variability of scores assigned to high-quality translation papers, a paper-by-rater (p x r) random effects G-study was carried out. Table 3 displays the components of variance and their relative contribution to the ratings assigned to high-quality translation papers.

Table 3. Variance components for random effects P X R design (high-quality translation papers)

Variance source	df	σ^2	%
P	39	0.38	13.1
R	9	1.72	58.5
PR	351	0.83	28.4
Total	399	-	100

Table 3 shows that the greatest variance component was attributable to raters (58.5%), indicating that 80 L2 translation papers were substantially different in terms of quality. The second largest variance component was determined to be the residual (28.4%). Papers yielded the smallest variance component (13.1%), indicating that the papers are somewhat different in terms of translation quality.

A paper-by-rater (p x r) random effects G-study was carried out to identify the sources of variance contributing to the scores assigned to low-quality translation papers. Table 4 shows the components of variance and their relative contribution to the variability of scores assigned to low-quality translation papers.

Table 4. Variance components for random effects P X R design (low-quality translation papers)

Variance source	df	σ^2	%
P	39	0.17	6.2
R	9	1.63	59.1
PR	351	0.95	34.7
Total	399	-	100

Table 4 shows that the greatest variance component was due to raters (59.1%), indicating that raters differed largely from one to another in terms of scoring the students' translations. The second largest variance component followed by the rater facet was the residual (34.7%). Papers yielded the smallest variance component (6.2%).

Calculations of generalizability and dependability coefficients were performed with the application of a paper-by-rater-by-quality (p x r x q) random effects G-study design for all translation papers, and a paper-by-rater (p x r) random effects G-study design for low- and high-quality translation papers separately. Table 5 illustrates these results.

Table 5. Generalizability and dependability coefficients for translation paper ratings

Translation papers	Number of papers	Number of raters	Ep^2	ϕ
All papers	80	10	.96	.95
High-quality	40	10	.82	.60
Low-quality	40	10	.64	.40

Table 5 shows that the generalizability and dependability coefficients of all translation papers (Ep^2 =.96 and ϕ=.95) were higher than those of high- and low-quality translation papers (Ep^2 =.82, ϕ=.60 and Ep^2 =.64, ϕ=.40, respectively). The results revealed that although Ep^2 and ϕ coefficients obtained for high-quality translation papers were higher than those obtained for low-quality ones, coefficients obtained for both quality of papers were far below the ones obtained for all papers.

6.2. Results for the second research question

Are there any significant differences among the scores assigned to low- and high-quality translation papers?

Mann-Whitney U test results for differences between low- and high-quality translation papers are given in Table 6.

Table 6. Mann-Whitney U test results for differences between low- and high-quality translations

Categories	U	p	High Quality (Mdn)	Low Quality (Mdn)
Purpose total	9383.000	0.000	0.8	0.4
Culture total	7628.000	0.000	1.6	0.6
Text total	8624.500	0.000	1.2	0.6
Technical aspects total	24614.500	0.000	1.2	0.7
Content total	7767.500	0.000	1.3	0.5
Lang total	7673.000	0.000	2.0	0.8
Grand total	6875.000	0.000	8.4	3.8

According to Table 6, the total purpose median was 0.8 in high quality texts, and 0.4 in low quality texts with statistically significant differences ($U=9383.000; p<0.05$).

The Purpose scores of high-quality texts were significantly higher than low-quality texts. The Culture score median was 1.6 in high-quality texts, and 0.6 in low-quality texts with statistically significant differences (U=7628.000; p<0.05). The Text score median was 1.2 in high-quality texts, and 0.6 in low-quality texts with statistically significant differences (U=8624.500; p<0.05). The Technical aspects score median was 1.2 in high-quality texts, and 0.7 in low-quality texts with statistically significant differences (U=24614.500; p<0.05). The Content score median was 1.3 in high-quality texts, and 0.5 in low-quality texts with statistically significant differences (U=7767.500; p<0.05). The Language score median was 2.0 in high-quality texts, and 0.8 in low-quality texts with statistically significant differences (U=7673.000; p<0.05). In total, the median of high-quality texts was 8.4, higher than low-quality texts(3.8) with statistically significant differences(U=6875.000; p<0.05). Thus, it might be argued that the evaluation of low-quality texts had higher variation than high-quality translation papers.

6.3. Results for the third research question

Does rating experience have an impact on the variability and reliability of the scores assigned to high- and low-quality translation papers?

The scores assigned to each component of the rubric were analysed to understand whether there were statistically significant differences between the scorings of experienced and inexperienced raters. Mann-Whitney U test results for differences between experienced and inexperienced raters were given in Table 7.

Table 7. Mann-Whitney U test results for differences between more experienced and less experienced raters

Categories	U	p	More Experienced (Mdn)	Less Experienced (Mdn)
Purpose total	77563.500	0.452	0.60	0.60
Culture total	71509.000	0.009	0.50	0.60
Text total	72406.500	0.019	0.30	0.30
Technical aspects total	71128.500	0.006	0.40	0.30
Content total	69638.000	0.001	0.30	0.40
Lang total	69479.500	0.001	0.30	0.30
Grand total	70131.000	0.003	1.30	1.60

According to Table 7, the total score median was 1.3 in more experienced raters, and 1.6 in less experienced raters with statistically significant differences (U=70131.000; p<0.05). In general, results showed that less experienced raters assigned higher scores to all papers than more experienced raters. Also, calculations of generalizability coefficients were performed for high- and low-

quality translation papers to determine differences between less experienced and more experienced raters. The results are given in Table 8.

Table 8. G-theory analysis results for differences between more experienced and less experienced raters

Translation papers	Number of papers	Number of raters	Ep^2 for more experienced	Ep^2 for less experienced	ϕ for more experienced	ϕ for less experienced
High-quality	40	5	.46	.80	.25	.52
Low-quality	40	5	.13	.55	.05	.32

As seen in Table 8, G-study analysis yielded higher Ep^2 and ϕ coefficients for both high- and low-quality translation papers (Ep^2: .80 and .55, and ϕ: .52 and .32, respectively) rated by less experienced raters. These results show that less experienced raters were more consistent in the scores they assigned to the papers than more experienced raters.

6.4. Results for the fourth research question

How do raters make decisions while assigning scores to translation papers of different quality?

Data obtained from the TAPs that were recorded by eight graders (2 of the raters failed to complete the TAPs task during their assessment) were analysed. In the presentation of the qualitative data, a cumulative approach was followed regarding translation quality. The most commonly employed strategies by all raters were identified for each translation quality. Table 9 shows the ranking orders of the decision-making behaviours.

Table 9. The most frequently employed decision-making behaviours by all raters to high- and low-quality translation papers

High-quality Translation Papers		**Low-quality Translation Papers**	
Decision-making Behaviours	%	Decision-making Behaviours	%
Read or reread text	32,60	Read or reread text	28,24
Consider spelling or punctuation	6,45	Consider spelling or punctuation	5,83
Articulate general impression	4,95	Assess comprehensibility	5,83
Consider syntax or morphology	4,72	Consider syntax or morphology	5,48
Consider own personal response, expectations, or biases	4,26	Interpret ambiguous or unclear phrases	4,46
Identify redundancies	3,46	Assess coherence	4,42
Read or interpret scoring scale	3,40	Consider own personal response, expectations, or biases	4,42
Assess comprehensibility	3,23	Envision personal situation of writer	3,71
Assess coherence	3,05	Summarize judgements collectively	3,36

Assess task completion or relevance	2,71	Identify redundancies	2,92
Assess style, register, or genre	2,71	Read or interpret scoring scale	2,61
Consider lexis	2,59	Consider lexis	2,61
Articulate or revise scoring	2,36	Assess task completion or relevance	2,34
Edit phrases for interpretation	2,36	Assess reasoning, logic, or topic development	2,12
Assess text organization	2,30	Articulate general impression	2,03
Assess reasoning, logic, or topic development	2,30	Rate language overall	1,94
Summarize judgements collectively	2,19	Edit phrases for interpretation	1,94
Interpret ambiguous or unclear phrases	2,07	Assess text organization	1,94
Rate language overall	1,96	Assess originality	1,86
Define or revise own criteria	1,73	Assess style, register, or genre	1,72
Assess originality	1,67	Scan or skim text	1,68
Scan or skim text	1,61	Consider gravity of errors	1,41
Consider gravity of errors	1,44	Classify errors into types	1,37
Assess fluency	1,32	Consider error frequency	1,37
Consider error frequency	0,92	Articulate or revise scoring	1,28
Classify errors into types	0,52	Assess fluency	1,24
Envision personal situation of writer	0,46	Compare with other compositions	0,75
Observe layout	0,35	Define or revise own criteria	0,71
Compare with other compositions	0,29	Rate ideas or rhetoric	0,27
Summarize ideas or propositions	0,00	Discern rhetorical structure	0,09
Rate ideas or rhetoric	0,00	Summarize ideas or propositions	0,04
Discern rhetorical structure	0,00	Assess quantity	0,00
Assess quantity	0,00	Decide on macro-strategy	0,00
Read prompt	0,00	Read prompt	0,00
Decide on macro-strategy	0,00	Observe layout	0,00
Total	100	Total	100

As seen in Table 9, across the two quality translation papers, seven of the top ten decision-making behaviours (with a different ranking order) employed by the raters were the same. All of the raters employed the same top two strategies ("read or reread text" and "consider spelling or punctuation") when assessing both the low- and high-quality translation papers and seemed to focus on language use by relying on the strategies of "consider syntax or morphology" and "consider spelling or punctuation". "Consider syntax or morphology" was another commonly preferred strategy with a ranking of 4th for both qualities of papers. Although the strategy of "assess comprehensibility" ranked 3rd for low-quality

translation papers, it ranked 8th for high-quality translation papers. Similarly, while the strategy of "assess coherence" ranked 6th for low-quality translation papers, it ranked 9th for high-quality translation papers. While the strategies of "consider own personal response, expectations or biases" and "identify redundancies" ranked 5th and 6th accordingly for high-quality translation papers, they ranked 7th and 10th respectively for low-quality translation papers.

Other strategies that appeared among the top ten most frequently used for high-quality translation papers were "articulate general impression" ranked 3rd, "read or interpret scoring scale", ranked 7th and "assess task completion or relevance" ranked 10th, but these strategies were ranked 15th, 11th and 13th respectively for low-quality translation papers. Similarly, the strategies "interpret ambiguous or unclear phrases", "envision personal situation of the writer," and "summarize judgments collectively" were more frequently employed for low-quality translation papers than high-quality translation papers. Although these strategies appeared in the top ten for low-quality translation papers, they were ranked 18th, 27th, and 17th respectively for high-quality translation papers. Considered collectively, these decision-making trends suggest that when assigning scores to low-quality translation papers, raters mostly centred on grammar, coherence, and comprehensibility of the texts. However, they focused more on identifying redundancies and on the general impression of the texts while rating high-quality translation papers.

7. Discussion

The first research question examined sources of score variation contributing to the score variability of the holistic scores assigned to high- and low-quality translation papers. Results revealed that the largest variance (61.8%) was attributable to papers (p). This indicates that students, as predicted, displayed different translation performances as measured by the translation task. On the other hand, when the generalizability analyses were performed on high- and low-quality translation papers, the facet of papers explained a relatively small portion of variance for low-quality papers in comparison to the variance observed in high-quality papers. Since there were more homogeneous student groups in each of the designs (high- and low-quality), variance due to the papers facet was expected to be small in these two groups. In this context, it can be said that students in the higher proficiency group performed more dissimilar translation abilities from each other when compared to the ones in the lower proficiency group. Also, this might have been due to the fact that some raters assigned lower scores to some of the high-quality papers than they deserved.

Generalizability analysis in the p x r x q design revealed that the second largest variance was due to the interaction between papers and raters (12.6%), indicating that some raters were inconsistent in their judgments while assigning scores to

certain translation papers. Following the interaction between raters and papers, the third greatest variance component in the same design was the residual (12.1%). This amount was relatively smaller when compared to the residual facet scorings in high-quality and low-quality translation papers (28.4% and 34.7%, respectively). These results indicate that some other factors such as gender, expectations, biases, methods of scoring, rating experience, educational background of the raters, etc., might have contributed to the score variation (Brennan, 2001; Huang et al., 2014; Şahan, 2018). Since these measurement designs included a limited number of facets, the residual facet was expected to have a high impact on the score variability.

Considering the impact of the rater facet on holistic score variability in three measurement designs, raters performed similar trends in terms of severity and leniency when assigning scores to high- and low-quality papers (58.5% and 59.1%, respectively). However, in the p x r x q design (when quality is included), the rater facet was determined to have no impact on the score variability (0.0%). These results show that raters follow more consistent rating trends while assigning scores to papers with different proficiency levels than while assigning scores to homogeneous groups of papers. That is to say, more consistent scores are likely to be obtained in large-scale assessment contexts where students' proficiency levels vary (Şahan, 2018). Regarding the other components of variance for the collective scorings of the papers, 12.1% of the total variance was attributable to the interaction between raters and paper quality. This revealed that raters varied considerably in the scores they assigned to high- and low-quality papers. These findings are parallel to what Şahan (2018) found in his thesis study. Furthermore, large variances were determined from the rater facet and the residual. This result might be related to the number of raters in that when the number of raters is increased, higher dependability coefficients are likely to be obtained (Brennan, 2001; Güler et al., 2012; Şahan, 2018).

The second research question was posited to determine whether there were any significant differences among the holistic scores assigned to low- and high-quality translations. Although the raters were not informed about the quality division in the translation paper pack, they could assign different scores to two different qualities of translation papers. However, the score range for the papers was found to be high, which might result from the contrast effect in that after assigning scores to a better or worse translation paper, the raters might show a tendency to rate another paper as higher or lower (Freedman, 1981). Furthermore, since raters were aware that EFL students had translated the texts, they may have had higher expectations regarding their performance. In light of this, the raters may have assigned lower scores to low-quality translation papers and to some of the high-quality translation papers. In a study by Baker (2010), raters were found to distinguish between high- and low-quality papers. However, they showed a tendency to give different scores to the same papers under different conditions

(either authentic or research conditions). Furthermore, the raters were found to be more consistent while scoring high-quality translation papers, while the variation was greater in the low-quality translation papers. While this finding coincides with the results of some of the previously held research (Han, 2017; Huang et al., 2014), it contradicts the findings of a study conducted by Şahan (2018), in which the raters were determined to be more consistent while assigning scores to low-quality papers. These results suggest that the quality of the paper is a variable impacting the score reliability. However, variation of scores between and among raters based on the quality of the paper might yield contradictory results in accordance with the raters' background and study context.

The third research question investigated the impact of the rating experience on the variability and reliability of the scores assigned to high- and low-quality translation papers. When previous rating experience was considered, it was determined that more experienced raters assigned lower scores to the translation papers than the less experienced raters. This finding indicates that while less experienced raters assessed the translation papers more leniently, more experienced raters assessed them more severely. The findings of this study coincide with the results of some previous research in that inexperienced raters tend to assign higher scores (Rinnert & Kobayashi, 2001), and they assess papers more leniently (Barkaoui, 2011; Sweedler-Brown, 1985). However, some of the previous research indicated that although raters with different rating experiences performed more similar analytic scorings, more experienced raters tended to assign higher holistic scorings (Song & Caruso, 1996). Also, some additional studies have found that more experienced raters tend to be more lenient than inexperienced raters when assigning scores to students' papers (Şahan, 2018; Weigle, 1999). However, in a study carried out by Shirazi (2019), both experienced and novice raters were determined to perform alike in terms of leniency and severity. Thus, previous research suggests that the way raters interpret the given rubric and to which criteria they afford priority may be a determinant of the dissimilarity of the ratings. In this study, low- and high-quality papers scored by less experienced raters yielded higher reliability coefficients than more experienced raters. These results suggest that in the scoring of both high- and low-quality translation papers, less experienced raters were more consistent than more experienced raters. This might be due to the fact that more experienced raters complied less with the criteria specified in the scoring scale and were more reliant on their own expectations (Eckes, 2008).

The final research question focused on identifying raters' decision-making behaviours while assigning scores to different quality of translation papers holistically. The three most commonly used strategies for all translation papers were "read or reread text," followed by "consider spelling or punctuation" and "consider syntax or morphology." The "Read or reread text" strategy was

expected to be employed frequently by the raters since they all were required to read both the source text and the translated versions of it at least once. This finding coincides with the findings of the previously held research (Barkaoui, 2010; Şahan, 2018). The fact that the writing task was translation may have caused the raters to concentrate on the appropriate use of syntax and morphology in the translation papers and whether the spelling and punctuation were in accordance with the target language. When decision-making behaviours used for low- and high-quality translation papers were compared, similarities were observed in the rating strategies of the raters. Across the two quality of translation papers, seven of the top ten decision-making behaviours (with a different ranking order) employed by the raters were the same.

In conclusion, the employed statistical analyses showed that raters differed substantially in the scores they gave to low-quality and high-quality translation papers, indicating that they could distinguish between low-proficient and high-proficient translation papers. Also, when compared to the high-quality translation papers, more variance was determined in the median scores of the low-quality translation papers, suggesting that raters were more consistent while rating high-quality translation papers. Considering the impact of the rater facet on holistic score variability in three measurement designs, raters performed similarly in terms of severity and leniency when assigning scores to high- and low-quality papers. However, the rater facet was determined to have no collective impact on the score variability. This revealed that raters varied considerably in the scores they assigned to high- and low-quality papers, and they displayed great differences regarding severity and leniency within each translation paper quality. Regarding the impact of previous experience on the reliability and variability of scores, it was determined that more experienced raters assigned considerably lower scores to the translation papers than the less experienced raters. Regarding the use of decision-making strategies, raters were found to apply mostly strategies pertaining to self-monitoring focus, followed by language focus and rhetorical/ideational focus, respectively.

This study is not without limitations. First, the lack of a thorough rater training procedure may have caused variations in scores among raters. Although necessary information was provided regarding the criteria included in the holistic rubric prior to the scoring process, it was observed that the rater training provided was insufficient to obtain fairer ratings. Second, performing verbal protocols on a translation assessment task might have led raters to be biased while making decisions about the papers (Şahan, 2018). Additionally, raters might have experienced pressure while thinking aloud, which might have introduced variables in regard to the quality of the verbal protocols (Barkaoui, 2010).

8. Conclusion

In light of the limitations and findings of this research, some pedagogical implications are suggested. First, although scoring training was provided to the raters prior to the assessment procedure of the students' translation papers, variations between the scorings of the raters were significant. For this reason, this study illustrates that even raters with a broad experience of rating should be given detailed and consistent rater training to make them more reliable markers. In this way, variations between the scorings can be considerably reduced or eliminated, and more fair judgments can be attained. Secondly, it is suggested that in-house scoring protocols and thorough rater training may be helpful for instructors to achieve more fair judgments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angelelli, C. V. (2009). *Testing and assessment in translation and interpreting studies.* John Benjamins.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191–208. https://doi.org/10.1016/j.jslw.2009.05.003

Baker, A. B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133–153. https://doi.org/10.1016/j.asw.2010.06.002

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2010). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75. https://doi.org/10.1177/0265532210376379

Berggren, I. (1972). *Does the use of translation exercises have negative effects on the learning of a second language?* Gothenburg University.

Brennan, R. L. (2001). *Generalizability theory.* Springer.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(4), 587–603. https://doi.org/10.2307/3587078

Brown, G. T. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

Calis, E., & Dikilitas, K. (2012). The use of translation in EFL classes as L2 learning practice. *Procedia-Social and Behavioral Sciences*, *46*, 5079–5084. https://doi.org/10.1016/j.sbspro.2012.06.389

Chang, Y. (2002). EFL teachers' responses to L2 writing. *Reports Research* (143). Retrieved from http://files.eric.ed.gov/fulltext/ED465283.pdf

Colina, S. (2008). Translation quality evaluation: Empirical evidence for a functionalist approach. *The Translator*, *14*(1), 97–134. https://doi.org/10.1080/13556509.2008.10799251

Cook, G. (2010). *Translation in language teaching: An argument for reassessment.* Oxford University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Wiley. https://doi.org/10.1126/science.178.4067.1275

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Dickins, J., Hervey, S., & Higgins, I. (2016). *Thinking Arabic translation: A course in translation method: Arabic to English*. Taylor & Francis.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185. https://doi.org/10.1177/0265532207086780

El-Banna, A. I. (1993). The development and validation of a multiple-choice translation test for ESL college freshmen. https://eric.ed.gov/?id=ED374661

Eyckmans, J., Anckaert, P., & Segers, W. (2009). The perks of norm-referenced translation evaluation, In C. V Angelelli, & H. E Jacobson, (Eds.), *Testing and assessment in translation and interpreting* (pp. 73–93). John Benjamins.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414–420. https://doi.org/10.2307/3587446

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245–255. https://www.jstor.org/stable/40170792

Ghaiyoomian, H., & Zarei, G. R. (2015). The effect of using translation on learning grammatical structures: A case study of Iranian junior high school students. *Research in English Language Pedagogy*, *3*(1), 32–39. https://relp.isfahan.iau.ir/article_533621.html

Ghonsooly, B. (1993): Development and validation of a translation test. *Edinburgh Working Papers in Applied Linguistics, 4*, 54–62. https://eric.ed.gov/?id=ED360838

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik kuramı*. Pegem Akademi Yayınları.

Han, C., & Shang, X. (2023). An item-based, Rasch-calibrated approach to assessing translation quality. *Target*, *35*(1), 63–96. https://doi.org/10.1075/target.20052.han

Han, T. (2017). Scores assigned by inexpert EFL raters to different quality of EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136–152. https://ijpe.inased.org/makale/229

Hatim, B., & Mason, I. (1997). *The translator as communicator.* Routledge.

Heaton, J. B. (2003). *Writing English language tests.* Longman.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87–107. https://doi.org/10.2307/3586337

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing*, *13*(3), 201–218. https://doi.org/10.1016/j.asw.2008.10.002

Huang, J., Han, T., Tavano, H., & Hairston, L. (2014). Using generalizability theory to examine the impact of essay quality on rating variability and reliability of ESOL writing. In J. Huang & T. Han (Eds.), *Empirical quantitative research in social sciences: Examining significant differences and relationships* (pp. 127–149). Untested Ideas Research Center.

Ito, A. (2004). Two types of translation tests: Their reliability and validity. *System*, *32*(3), 395–405. https://doi.org/10.1016/j.system.2004.04.004

Källkvist, M. (1998). How different are the results of translation tasks? A study of lexical errors. In K. Malmkjær (Ed.), *Translation and language teaching: Language teaching and translation* (pp. 77–87). St Jerome.

Källkvist, M. (2008). L1–L2 translation versus no translation. In L. Ortega, & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp.182–202). Routledge.

Lado, R. (1964). *Language teaching, a scientific approach.* McGraw-Hill.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716. https://doi.org/10.1093/applin/amn018

Laukkanen, J. (1996). Affective and attitudinal factors in translation processes. *International Journal of Translation Studies*, *8*(2), 257–274. https://doi.org/10.1075/target.8.2.04lau

Lee, T. Y. (2013). Incorporating translation into the language classroom and its potential impacts upon L2 learners. In D. Tsagari, & G. Floros (Eds.), *Translation in language teaching and assessment* (pp. 3–22). Cambridge Scholars Publishing.

Marais, K. (2013). Constructive alignment in translator education: reconsidering assessment for both industry and academy. *Certification, 5*(1), 13–31. https://search.informit.org/doi/abs/10.3316/informit.282441134346671

Melis, N. M., & Albir, A. H. (2001). Assessment in translation studies: Research needs. *Meta, 46*(2), 272–287. https://doi.org/10.7202/003624ar

Mundt, K., & Groves, M. (2016). A double-edged sword: the merits and the policy implications of Google Translate in higher education. *European Journal of Higher Education*, *6*(4), 387–401. https://doi.org/10.1080/21568235.2016.1172248

Neves, R. R. (2002). Translation quality assessment for research purposes: An empirical approach. *Cadernos de Tradução*, *2*(10), 113–131. https://doi.org/10.5007/%25x

Orozco, M. (2000). Building a measuring instrument for the acquisition of translation competence in trainee translators. *Benjamins Translation Library*, *38*, 199–214. https://doi.org/10.1075/btl.38.19oro

Pöchhacker, F. (1994). *Simultan dolmetschen als komplexes handeln.* Narr.

Prince, P. (1996). Second language vocabulary learning: The role of context versus translation as a function of proficiency. *The Modern Language Journal*, *80*(4), 478–493. https://doi.org/10.1111/j.1540-4781.1996.tb05468.x

Pym, A. (2010). *Exploring translation theories*. Routledge.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal, 85*(2), 189–209. https://doi.org/10.1111/0026-7902.00104

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. https://doi.org/10.1037/0003-066X.44.6.922

Shirazi, M. A. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open, 9*(1), 1–14. https://doi.org/10.1177/2158244018822377

Soleimani, H., & Heidarikia, H. (2017). The effect of translation as a noticing strategy on learning complex grammatical structures by EFL learners. *Applied Linguistics Research Journal*, *1*(1), 1–13. https://doi.org/10.14744/alrj.2017.09797

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5*(2), 163–182. https://doi.org/10.1016/S1060-3743(96)90023-5

Stapleton, P., & Kin, B. L. K. (2019). Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. *English for Specific Purposes*, *56*, 18–34. https://doi.org/10.1016/j.esp.2019.07.001

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal, 74*(5), 49–55. https://doi.org/10.2307/817702

Şahan, Ö. (2018). *The impact of rating experience and essay quality on rater behavior and scoring* [Unpublished doctoral dissertation]. Çanakkale Onsekiz Mart University.

Tavakoli, M., Shafiei, S., & Hatam, A. H. (2012). The relationship between translation tests and reading comprehension: A case of Iranian University students. *Iranian Journal of Applied Language Studies*, *4*(1), 193–211. https://doi.org/10.22111/IJALS.2012.1353

Uzawa, K. (1996). Second language learners' processes of L1 writing, L2 writing, and translation from L1 into L2. *Journal of Second Language Writing*, *5*(3), 271–294. https://doi.org/10.1016/S1060-3743(96)90005-3

Vaezi, S., & Mirzaei, M. (2007). The effect of using translation from L1 to L2 as a teaching technique on the improvement of EFL learners' linguistic accuracy–focus on form. *Humanising Language Teaching*, *9*(5), 79–121. http://old.hltmag.co.uk/sep07/mart03.rtf

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 717–719). Wiley.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2)*,* 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Yıldız, M. (2020). A critical perspective on the translation quality assessments of five translators` organizations: ATA, CTTIC, ITI, NAATI, and SATI. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi,* (18), 568–589. https://doi.org/10.29000/rumelide.706390