



**ÜNİVERSİTE ÖĞRENCİLERİNİN ÇEVİRİ
KALİTESİNİN DEĞERLENDİRİLMESİ
SÜRECİNİN İNCELENMESİ**

Nilüfer AYBİRDİ

Doktora Tezi

Yabancı Diller Eğitimi Ana Bilim Dalı

2020

(Her hakkı saklıdır.)

T.C.
ATATÜRK ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI
İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

**ÜNİVERSİTE ÖĞRENCİLERİNİN ÇEVİRİ KALİTESİNİN DEĞERLENDİRİLMESİ
SÜRECİNİN İNCELENMESİ**

(An Investigation into Evaluating Translation Quality of University Students)

DOKTORA TEZİ

Nilüfer AYBİRDİ

Danışman: Prof. Dr. Mehmet TAKKAÇ

Erzurum
Ağustos, 2020

KABUL VE ONAY TUTANAĞI

Nilüfer AYBİRDİ tarafından hazırlanan “Üniversite Öğrencilerinin Çeviri Kalitesinin Değerlendirilmesi Sürecinin İncelenmesi” başlıklı çalışması .. / .. / 20.. tarihinde yapılan tez savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yabancı Diller Eğitimi Ana Bilim Dalı, İngiliz Dili Eğitimi Bilim Dalında doktora tezi olarak kabul edilmiştir.

Jüri Başkanı:	Doç. Dr. Emrah EKMEKÇİ <i>Ondokuz Mayıs Üniversitesi</i>	Aslı Islak İmzalıdır
Danışman:	Prof. Dr. Mehmet TAKKAÇ <i>Atatürk Üniversitesi</i>	Aslı Islak İmzalıdır
Jüri Üyesi:	Dr. Öğr. Üyesi Oktay YAĞIZ <i>Atatürk Üniversitesi</i>	Aslı Islak İmzalıdır
Jüri Üyesi:	Dr. Öğr. Üyesi Oktay AKARSU <i>Atatürk Üniversitesi</i>	Aslı Islak İmzalıdır
Jüri Üyesi:	Dr. Öğr. Üyesi Özgür ŞAHAN <i>Yozgat Bozok Üniversitesi</i>	Aslı Islak İmzalıdır
İkinci Tez Danışmanı	Doç. Dr. Turgay HAN <i>Ordu Üniversitesi</i>	Enstitü Yönetim Kurulunun .../.../.... tarih ve sayılı kararı.

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım.

Prof. Dr. Mustafa SÖZBİLİR

Enstitü Müdürü

Aslı Islak İmzalıdır

ETİK VE BİLDİRİM SAYFASI

Doktora Tezi olarak sunduğum “Üniversite öğrencilerinin çeviri kalitesinin değerlendirilmesi sürecinin incelenmesi” başlıklı çalışmanın tarafımdan bilimsel etik ilkelere uyularak yazıldığını ve yararlandığım eserleri kaynakçada gösterdiğimi beyan ederim.

22 /07 / 2020

Nilüfer AYBİRDİ

Aslı Islak İmzalıdır

☐ Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun/.../.... tarih ve sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

☐ Enstitü Yönetim Kurulunun/.../.... tarih ve sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

ACKNOWLEDGEMENT

Writing a PhD dissertation is not an effortless process with the addition of many challenges. For this reason, it needs the contribution of not only the researcher but also many different people including the participants of the study. This dissertation would be impossible without the support of valuable people. That's why, I would like to express my sincere gratitude and thanks to my supervisors, colleagues, and participants.

I would like to express my deep gratitude to my supervisor Prof. Dr. Mehmet TAKKAÇ, who supported and guided me at every stage of this thesis study. Also, I would like to Express my sincere gratitude to my co-supervisor Assoc. Prof. Dr. Turgay HAN for the patience, guidance, support and constructive and helpful criticisms he provided throughout the entire process. In addition, I would like to heartily thank Inst. Catherine AKÇA for proofreading and editing this thesis study. Finally, I would like to thank 2nd year students of Kafkas University, Department of English Language and Literature, and instructors who participated in the study voluntarily for their invaluable contributions to the data collection process of the research.

Nilüfer AYBİRDİ

ÖZ

DOKTORA TEZİ

ÜNİVERSİTE ÖĞRENCİLERİNİN ÇEVİRİ KALİTESİNİN DEĞERLENDİRİLMESİ SÜRECİNİN İNCELENMESİ

Nilüfer AYBİRDİ

Ağustos 2020, 177 Sayfa

Amaç: Bu tezin amacı, çeviri kalitesi ve puanlama deneyiminin puanlayıcıların puanlama davranışları ve puanların değişkenliği üzerindeki etkisini incelemektir.

Yöntem: Karma yöntem araştırma deseninin uygulandığı çalışmanın verileri 40 ikinci sınıf YDİ üniversite öğrencisinden ve Türkiye'deki yüksek öğretim kurumlarında görev yapan 10 öğretim elemanından toplanmıştır. Bildirdikleri puanlama deneyimlerine göre, değerlendiriciler iki gruba ayrılmıştır. Öğrencilere uygulanan dört çeviri görevinin ardından, çalışmada kullanılmak üzere rastgele olarak 40 yüksek kaliteli ve 40 düşük kaliteli çeviri kağıdı seçilmiştir. Bu iki farklı kalitedeki kağıtlar bütünsel bir değerlendirme ölçeği kullanılarak puanlanması için değerlendiricilere verilmiştir ve çeviri kağıtlarını değerlendirirken karar verme davranışlarını anlamak amacıyla önceden belirlenmiş bazı çevirilere sesli düşünme protokolleri kaydetmeleri istenmiştir. Yüksek kaliteli ve düşük kaliteli çeviri kağıtlarına atanan puanların değişkenliği nicel olarak incelenirken, değerlendiricilerin karar verme davranışları nitel olarak incelenmiştir. Verilerin analizinde Genellenebilirlik (G-) teorisi yaklaşımı ile tanımlayıcı ve çıkarımsal istatistikler kullanılmıştır.

Bulgular: Çalışmanın bulguları, deneyimsiz değerlendiricilerin öğrencilerin çeviri kağıtlarına karşı daha olumlu olduğunu ve deneyimli değerlendiricilere kıyasla kağıtlara daha yüksek ve daha tutarlı puanlar verdiğini ortaya koymuştur. Ayrıca, G-teorisi analizleri, her bir kağıt kalitesi için ayrı analizler yapıldığında, puanlayıcı yüzeyinin puan değişkenliği üzerinde büyük bir etkisi olduğunu göstermiştir. Buna ek olarak, nitel veri analizi, değerlendiricilerin yüksek kaliteli kağıtlara kıyasla düşük kaliteli çeviri kağıtlarını değerlendirirken daha fazla dile ilişkin stratejiler kullandığını göstermiştir.

Sonuç: Çalışmanın bulguları çeviri değerlendirmesinde puanlayıcı güvenilirliği hakkında önemli veriler sağlamaktadır.

Anahtar Kelimeler: çeviri değerlendirmesi, çeviri kağıdı kalitesi, genellenebilirlik teorisi, puanlayıcı deneyimi, puan değişkenliği, sesli düşünme, karar verme stratejileri

ABSTRACT

DOCTORAL DISSERTATION

AN INVESTIGATION INTO EVALUATING TRANSLATION QUALITY OF UNIVERSITY STUDENTS

Nilüfer AYBİRDİ

August 2020, 177 Pages

Aim: The purpose of this study was to examine the impact of translation quality and rating experience on raters' scoring behaviors and on the variability of scores.

Method: Following a mixed-methods research design, the data of the study were collected from 40 second year EFL university students and 10 instructors working at higher education institutions in Turkey. Based on their reported rating experience, the raters were divided into two groups. After the administration of four translation tasks to the students, 40 high-quality and 40 low-quality translation papers were chosen randomly to be used in the study. The raters were asked to assign scores to these two distinct qualities of papers by using a holistic rubric and they were required to record think-aloud protocols for some pre-determined papers to understand their decision-making behaviors. The variability of the scores assigned to high-quality and low-quality translation papers was examined quantitatively while the decision-making behaviors of the raters were investigated qualitatively. Generalizability (G-) theory approach and descriptive and inferential statistics were performed for the analysis of the data.

Findings: The results of the study revealed that inexperienced raters were more positive toward students' translation papers and assigned higher and more consistent scores to papers when compared to experienced raters. Furthermore, G-theory analyses performed on each paper quality revealed that the rater facet had a great impact on score variability. In addition, qualitative data analysis showed that raters used more language-related strategies while assessing low-quality translation papers compared to high-quality papers.

Conclusion: In conclusion, the findings of this study provide significant insights into rater reliability in translation assessment.

Keywords: translation assessment, translation paper quality, generalizability theory, rater experience, score variability, thinking-aloud, decision-making strategies

TABLE OF CONTENTS

KABUL VE ONAY TUTANAĞI.....	i
ETİK VE BİLDİRİM SAYFASI.....	ii
ACKNOWLEDGEMENT.....	iii
ÖZ.....	iv
ABSTRACT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER ONE.....	1
Introduction	1
Place of Translation in Departments of Foreign Language Education in Turkey.....	1
Evaluation in Translation	2
Problem Statement	6
Purpose of the Study	7
Significance of the Study	8
Limitations of the Study.....	9
Definitions of Key Terms.....	10
Organization of Thesis	11
CHAPTER TWO.....	13
Review of Literature	13
Translation and Translation Quality.....	13
Assessment and Assessment Types	15
Subjectivity and Prescriptive Judgment in Translation Assessment.....	17
Reliability, Validity and Fairness Issues in Translation Assessment.....	19
Approaches to Translation Quality Assessment	20
Subjective (anecdotal) and mentalist approaches to evaluating translations.	20
Behavioral approaches to judging translation quality.	21
Functionalist approaches to assess translations.....	22
Major Translation Quality Assessment Models	22
Translation quality assessment model of Larose.	23

Williams' argumentation-centered assessment model.	24
Al-Qinai's translation quality assessment model.	25
House's translation quality assessment model.	26
Empirical Studies on Translation Quality Assessment	29
Comparative studies on translation assessment.	30
Studies on reliability and validity of translation tests.	36
Studies conducted on the use of translation as an assessment tool.	41
Studies on the impact of translation on L2 learners' accuracy.	44
Summary and Research Gaps in Translation Assessment	48
CHAPTER THREE	50
Methodology.....	50
Theoretical Framework	51
Dynamic equivalence theory.....	51
Constructive alignment: theoretical construct of the instrument used in the study.	53
Generalizability theory.....	54
Selection of Raters	60
Selection of Translation Students	62
Data Collection Instruments.....	62
Selection of Translation Passages	63
Readability of the articles.....	64
Characteristic features of the articles.	64
Rating Scale (Holistic Rubric) Used in the Study.....	65
Think-Aloud Protocols (TAPs).....	66
Combining introspection technique with TAPs.	68
Data Collection Procedure	68
Rating Procedure	69
Voice-recording Raters' Verbal Thoughts	69
Role of the Researcher	70
Validity and Reliability	70
Validity of the holistic rubric.	70
Single rater reliability.....	70
Inter-rater reliability.....	71
Preparation of the Data.....	71
Preparation of quantitative data.	71
Preparation of qualitative data.	72

Research Ethics	74
Summary	74
CHAPTER FOUR	75
Findings	75
CFA Analysis Results for Validity and Reliability of Rubric	75
Quantitative Data Analysis Results.....	76
Findings Related to Research Questions.....	77
Results for the first research question.	77
Results for the second research question.	80
Results for the third research question.	84
Results for the fourth research question.....	87
Results of Qualitative Data Analysis	92
Results for the fifth research question.....	93
Summary of the Findings	101
CHAPTER V	103
Discussion and Conclusion.....	103
Discussion	103
Discussion of findings for RQ1.....	103
Discussion of findings for RQ2.....	105
Discussion of findings for RQ3.....	106
Discussion of findings for RQ4.....	108
Discussion of findings for RQ5.....	111
Conclusion	114
Pedagogical Implications	116
Methodological Implications	117
Further Research	118
REFERENCES	121
APPENDICES	135
Appendix-A. Checklist for Holistic Rubric	135
Appendix-B. Rater Profile Form.....	136
Appendix-C. Consent Form for Students.....	137
Appendix-D. Holistic Scoring Rubric.....	138
Appendix-E. Source Texts Obtained from Daily Star Online Magazine.....	139
Appendix-F. TAP Evaluation Guidelines for Raters	143

Appendix-G. Official Permission Form Obtained from Kafkas University Faculty of Science and Letters	146
Appendix-H. Written Consent Form (for raters).....	153
Appendix-I. (p x r x q) Random Effects G-study for All Translation Papers.....	154
Appendix-J. (p x r) Random Effects G-study for High-Quality Translation Papers	155
Appendix-K. (p x r) Random Effects G-study for Low-Quality Translation Papers....	156
Appendix-L. Reliability and Dependability Coefficients of Low- and High-Quality Translation Papers by Gender	157
Appendix-M. Reliability and Dependability Coefficients of Low- and High-Quality Translation Papers by Experience	159
CURRICULUM VITAE	161



LIST OF TABLES

Table 1. <i>Summary of Reviewed Articles</i>	29
Table 2. <i>Rating Experience of the Participants</i>	60
Table 3. <i>Gender and Age Distribution of the Raters</i>	61
Table 4. <i>Participants' Academic Title and Institutional Affiliation Distribution</i>	61
Table 5. <i>Participants' Highest Level of Education and Previous Training on Writing Assessment</i>	61
Table 6. <i>Readability Analyzer Statistics of the Four Passages</i>	64
Table 7. <i>Techniques Related to Think Aloud Method</i>	67
Table 8. <i>CFA Results and Fitness Criteria</i>	76
Table 9. <i>Reliability Analysis Results</i>	76
Table 10. <i>Mann-Whitney U test Results for Differences Between Low- and High-Quality Translations</i>	77
Table 11. <i>Kruskal Wallis Test Results for Score Variability Between and Among Raters' Scores (Mdn)</i>	81
Table 12. <i>Kruskal Wallis Test Results for Score Variability Between and Among Raters' Scores (Mdn)</i>	82
Table 13. <i>Variance Components for Random Effects P X R X Q Design</i>	84
Table 14. <i>Variance Components for Random Effects P X R Design (High-quality Translation Papers)</i>	85
Table 15. <i>Variance Components for Random Effects P X R Design (Low-quality Translation Papers)</i>	85
Table 16. <i>Generalizability and Dependability Coefficients for Translation Paper Ratings</i>	86
Table 17. <i>Generalizability and Dependability Coefficients for All, High and Low Translation Papers</i>	87
Table 18. <i>Mann-Whitney U Test Results For Differences Between Female and Male Raters</i>	88
Table 19. <i>G-theory Analysis Results For Differences Between Female and Male Raters</i>	89
Table 20. <i>Mann-Whitney U Test Results for Differences Between Experienced and Inexperienced Raters</i>	90
Table 21. <i>G-theory Analysis Results For Differences Between Experienced and Inexperienced Raters</i>	91
Table 22. <i>Spearman's Rho Correlation Analysis Results For Task and Total Scores</i>	92
Table 23. <i>The Most Frequently Employed Decision-Making Behaviors by all Raters to all Translation Papers</i>	95

Table 24. *The Most Frequently Employed Decision-Making Behaviors by all Raters to High- and Low-Quality Translation Papers* 96

Table 25. *Summary of the Findings Belonging to Each Research Question*.....101



LIST OF FIGURES

<i>Figure 1.</i> House's scheme for analyzing and comparing texts (1997, p.108).....	28
<i>Figure 2.</i> House's scheme for translation errors (1997, p.108).....	28
<i>Figure 3.</i> Diagram of the applied convergent parallel design (Creswell & Plano Clark, 2015, p. 56).	51
<i>Figure 4.</i> Nida's three-stage system of translation (Nida & Taber, 1969, p. 33).....	53
<i>Figure 5.</i> Distribution of the score weights to the rubric's subscales.	66
<i>Figure 6.</i> The think aloud memory.....	67
<i>Figure 7.</i> CFA results for rubric.....	75
<i>Figure 8.</i> Distribution of grand total scores for high quality translation passages.....	79
<i>Figure 9.</i> Distribution of grand total scores for low quality translation passages.....	80
<i>Figure 10.</i> Differences between female and male raters regarding the scores.....	89
<i>Figure 11.</i> Differences between experienced and non-experienced raters regarding their scorings.	91
<i>Figure 12.</i> Distribution of strategy type in terms of quality of translation papers.	94
<i>Figure 13.</i> Distribution of strategy focus in terms of quality of translation papers.	94

LIST OF ABBREVIATIONS

CFA	: Confirmatory Factor Analysis
D-study	: Decision Study
EFL	: English as a Foreign Language
ESL	: English as a Second Language
FL	: Foreign Language
G-study	: Generalizability Study
G-theory	: Generalizability Theory
L1	: First Language
L2	: Second Language
NoG	: No Grammar Group
NoT	: No Translation Group
T	: Translation Group
TAP	: Think Aloud Protocol
TQA	: Translation Quality Assessment

CHAPTER ONE

Introduction

Place of Translation in Departments of Foreign Language Education in Turkey

In our universities, translation education is regarded as an effective teaching / learning technique in both undergraduate and graduate programs related to foreign language teaching and as a professionally applied language teaching approach (Sönmez, 1997, p.33). Translation is not only a technique used in the programs aimed at educating translators in our country, but it also covers all the translation courses in the curricula of foreign language teaching programs (Sönmez, 1997, p.33). Moreover, many private and state universities in Turkey aim at providing students with academic skills such as preparing projects, listening, speaking and writing in programs whose language of education is English (e.g., medicine, business) (Gündüz, 2016). In this method, which is known as English for specific purposes, translation is used as the basic methodology in addition to the existing language activities, (Sönmez, 1997, p.33). It is obvious that translation is used as a pedagogical tool in these courses for the purpose of improving students' abilities to use written information in the foreign language.

In Turkey, in departments of Foreign Language Education, translation courses are included in the curricula in order to improve students' translation skills. Although the final aim of these departments is to train foreign language teachers, almost half of the students who graduate from the related departments ultimately work in fields other than teaching (Köksal, 2005, p.120). Although the employment expectations of the students studying in the departments of Foreign Language Education vary, translation is the most important skill that students will need in the areas they plan to study (Köksal, 2005, p.120). Currently, in the departments of Foreign Language Education of the Faculties of Education and in the English Language and Literature departments of the Faculties of Literature, from first-grade to last-grade, courses in translation, from the mother tongue to the foreign language, from the foreign language to the mother tongue, and translation of literary texts, are taught at all levels although the levels vary according to the departments' courses.

In Turkish universities, among many other types of performance assessment, assessing translation performance is mostly performed for the purposes of evaluating students' linguistic progress and making them more familiar with the culture of the target language to which they are exposed. Since each of these purposes impacts learners' improvement in higher education,

it is necessary that the written products of the students are assessed using a reliable method. Due to the aforementioned purposes, instructors who are responsible for giving translation courses in such institutions mostly prepare the exams and generally do not employ any formal procedures while evaluating the papers. Also, most of the instructors do not have any previous rater training experience; thus, they usually apply their own inner criteria and perceptions when assigning scores to the students' papers. While some instructors use somewhat reliable rubrics for scoring, some of them use anonymous evaluation procedures. In most cases, translation assessment is carried out by a single rater and the transparency of the students' identities affects the process of assessment. Furthermore, the employment of different grading methods at different institutions, and even within the same institution, leads to biased evaluation in Turkish universities. Therefore, this situation highlights the urgent need for a sound and standardized grading system to prevent unfair judgment.

Evaluation in Translation

In our world, which is gradually shrinking with the development of science and technology, the phenomenon of translation retains its importance and relevance, along with all its problems, as an inseparable part of communication both at the linguistic and inter-linguistic level (Aktaş, 1996). One of the main issues that researchers in the field of translation studies agree on is that the act of translation is a complex multidimensional process. Translation can be defined as a transfer between languages and cultures (Boztaş, 1992); as an activity that bridges various civilizations, bringing together individuals from different societies; and as a means of communication that transfers the historical, traditional and social elements of a civilization to other civilizations (Aktaş, 1996). In this sense, it can be said that the act of translation is a challenging task requiring rigor.

Since language depicts the way that the people of a particular language perceive the world, it is part of a culture which is peculiar to a group of people or society and makes them different from others. In translation, two different cultures peculiar to two different societies are confronted (Ateşman, 2001). But, how do we determine if a translation is good enough? This simple question lies at the center of all issues related to translation evaluation (House, 2001). In general, translation evaluation is defined as the value judgments of an expert (evaluator), based on intuition and experience rather than experimentally demonstrable data (Martinez Mateo, 2016). Aksoy (2001) argues that in order to be able to evaluate a translation as put forward in translation studies, it is necessary to determine carefully the limits and requirements of this effort. It is essential for the evaluator to know what the translation effort requires, what the task of the translator is, and how the problems of translation can be overcome.

Also, the evaluator should examine the source text in a systematic way by emphasizing language, culture, message and literary features. Martinez Mateo (2016) defines three important stages to be followed in measuring translation quality; 1) defining what value / quality is, 2) determining the methodology to be applied, 3) carrying out the actual evaluation procedure according to the definition of pre-determined methodology and quality.

However, House (2001) claims that the theoretical bases of the judgments reached in translation evaluation should be sound and interpersonally negotiable with certain verifiable procedures. According to the first translation theories, translation evaluation was carried out in a non-systematic way in relation to a single criterion (Nida, 1964). Besides, evaluations were mostly based on binary criteria, such as word for word or sense for sense translation (Medadian & Mahabadi, 2015). In the past, translation was a frequently applied method for teaching languages, until it was abandoned with the emergence of communicative teaching methods. However, recently many theorists and educationalists have approached translation as a profitable communicative learning activity. Before the 1980s, all approaches to translation theory dealt with translation as competence and with the translated text as product. Translation competence of a translator was used to be analyzed by comparing the translated text with the source text according to the criteria of various equivalence models. In the 1980s, some theorists attempted to propose more objective translation models for translation evaluation. However, since they were all based upon idealizations, these early translation models lack practical applicability and could not be tested empirically (Toury, 1995). Also, it is difficult to assess inter-rater reliability and replicability in these approaches, since they mostly lack explicitly defined criteria for translation assessment. Thus, the results show significant variation, with a high level of subjectivity amongst raters.

In the 1990s and early 2000s, some scholars claimed that there was a scarcity of empirical research in the field of translation studies (e.g., Hatim & Mason, 1997; Melis & Albir, 2001; Pym, 1992). They asserted that although a few studies had been carried out on translation assessment, they had been conducted neither objectively nor in a disciplined fashion (Melis & Albir, 2001, p. 273). A theoretical and descriptive approach is taken to the issue of translation assessment even in the latest publications. This situation clearly exposes the absence of empirical studies in the field of translation quality assessment (TQA) (Waddington, 2001, p. 18). From the 1990s to the present time, the amount of empirical research carried out in the field of TQA is limited. Moreover, studies searching for an objective method to assess translated works are even scarcer. These studies have focused primarily on the examination of the reliability and validity of developed translation tests (e.g., Colina, 2008; El-Banna, 1993;

Eyckman, Anckaert & Segers, 2009; Ghonsooly, 1993; Ito, 2004; Neves, 2002; Orozco, 2000; Tavakoli, Shafiei & Hatam, 2012).

Since translation testing methodology has always been considered to have a subjective nature, in his study Ghonsooly (1993) compared a translation test with a standardized test in terms of its validity and reliability as an assessment tool. In the adaptation process of the translation test, the researcher followed some procedures. Two phases of pretesting were performed in the study: in the first phase, a preliminary version of the test, containing 40 items in an open-ended form was administered to 825 volunteer engineering students; in the second phase, the same test was given to two translation experts. Then, the students' answers were compared with the experts' in order to constitute a multiple-choice translation test and to ensure its objectivity. The final version of the translation test was administered to 60 volunteer engineering students. The reliability of the test was determined as .74. The concurrent validity of the test was found to be low and not significant. Factor analysis of the translation test showed that it could be used to measure students' comprehension skills and some other basic language skills such as grammar and vocabulary.

Again in 1993, El-Banna carried out a quasi-experimental research project in order to develop a multiple-choice translation test that could be used as an assessment tool in determining the level of proficiency of learners of English as a Second Language (ESL) in translating texts. The pilot test was scored by the researcher and the final version of the test (composed of two parts, each including 30 items) was constructed. Then, the test was given to another 92 freshmen college students. The intrinsic validity of the test was found to be .82. The reliability coefficient of the first part of the test was determined as .66 and second part's as .67, indicating that the multiple-choice translation test was good at measuring learners' translation skill and proficiency levels. In 2000, Orozco conducted a research project in order to assess first year translation students' acquisition of translation competence. First, dependent and independent variables (translation errors, translation problems and translation notions) were tested by three distinct experimental tools. Afterwards, validation and standardization processes were carried out for each test and all of them were found to be reliable. Each of the tests was administered to translation students from three state universities at two different times (one at the beginning of the first semester and the other after an eight-month of interval). Analysis of the translations revealed four major problems: (1) grammatical problem, (2) extra-linguistic problem, (3) transfer problem, and (4) pragmatic problem.

In 2002, Neves carried out an experiment in order to examine the reliability of a quality scale developed for translation assessment. For quality assessment, six undergraduate students

attending translation courses in Brazil and six professional translators were asked to translate an English text into Portuguese. As a result of the study, the scale was determined to be reliable with .95 Cronbach Alpha coefficient. However, according to the evaluators, the scale was not sufficient enough for measuring the construct. Regarding the consensus between the scores allocated by each rater to each question in the scale, highly significant correlations were determined. In 2004, Ito compared open-ended translation tests with multiple-choice translation tests in terms of their effectiveness in improving learners' reading comprehension skills. Also, the researcher attempted to determine whether the language proficiency level of the students had an impact on the validity and reliability of the tests. In total, 35 male and 35 female Japanese high school students with similar language backgrounds and educational levels participated in this experimental study. The completed test papers were rated by two Japanese English language teachers from a state university. The inter-rater reliability coefficients were measured as between .83 and .90. The findings of the study indicated that the proficiency level of the students had a significant impact on the reliability and validity of the tests.

Colina (2008) developed a translation quality assessment tool centered upon a functionalist/componential approach. In order to provide some empirical evidence for the proposed approach and to measure the reliability of the tool, the researcher conducted a testing study. A total of 40 raters with three different first language (L1) backgrounds (Chinese, Spanish or Russian) participated in the study. Before the beginning of the rating procedure, the raters were given information about the tool and they were trained in how to use it. The results of the study indicated good inter-rater reliability: by text .8 and by language groups .7. In the following year, Eyckman, *et al.* (2009) compared a norm-referenced method (The Calibration of Dichotomous Items) that they developed for translation assessment with two other methods (analytical and holistic) in terms of their reliability. The participants of the study were 113 university students attending Dutch-French translation courses at four different francophone translation colleges in Belgium. Four professional raters (two male and two female) were recruited for scoring the translations of the students according to the method they preferred to use. The reliability of the test was calculated and it was found to have .958 Cronbach Alpha Coefficient. On the other hand, the correlation coefficients between the graders ranged between .635 and .780. The results of holistic assessment revealed less inter-rater reliability than those of analytic grids. Lastly, Tavakoli *et al.* (2012) attempted to examine the reliability and validity of translation tests, seeking to prove that translation could be applied as a reliable assessment method for measuring the reading comprehension level of learners. To achieve this objective, initially, the researchers compared an open-ended translation test with a multiple-choice translation test. Then, these two tests were contrasted with a multiple choice reading

comprehension test and an open-ended cloze test. A number of statistical tests were performed and the open-ended translation test's reliability coefficient was found to be .87 and its validity .89. The multiple-choice translation test was found to be reliable (.79) and moderately valid (.69). Therefore, the open-ended translation test was defined as more reliable and valid than the multiple-choice translation test.

Today, when the quality of translation is evaluated professionally, intuitive and analytical procedures are followed (Martinez-Mateo, 2016). While the intuitive translation evaluation method is based on personal judgments, the analytical method evaluates translation with a systematic approach (Martinez-Mateo, 2016).

Problem Statement

According to Lado (1964), the presence of human (teacher) interference in the process of scoring translation tests gives it a highly subjective character that is indicative of unreliability. When the evaluation of a translation is carried out by a person (teacher), many factors, such as personal thoughts, beliefs, assumptions and prejudices affect the process of translation evaluation, causing a rather subjective and inconsistent assessment. Just as the performance of an individual in a test can be affected by psychological and physical factors such as stress, anxiety, fatigue, motivation and noise, the evaluation of a performance might be affected by the rater having difficulty in determining a consistent scoring criterion or by different scoring methods followed by different raters (Eyckmans & Anckaert, 2017). When evaluating the outcome of a translation procedure, the social, political and even the ethical stance of the evaluator might have an impact on the evaluation process (Rabadan, Labrador, & Ramón, 2009). Although the intuition of experts is considered to be reasonably reliable, field researchers still need measurement tools that are formulated in an experimentally validated model and more reliable procedures to make the evaluation process more effective and more objective (Al-Melhi, 2014).

In the case of translation evaluation, a student's translation of a text from the source language into the target language is seen as a valid way to verify whether the student can actually translate between these two languages. Although the majority of translation examinations involve the task of translating texts and although this task is accepted as a valid translation adequacy measure, it seems that the validity discussions that take place in translation evaluation literature are inadequate (Eyckmans & Anckaert, 2017). In assessing structure-independent variables specific to the holistic and analytical scoring, scientists have abandoned the search for "objective" assessment in recent years, focusing on inter-rater reliability, intra-

rater reliability, and the construct validity of translation tests. (Eyckmans & Anckaert, 2017; Waddington, 2001).

Thus, in order to better understand the issue of variability of rating, it is essential to have an insight into commonalities and differences in rater reactions to translations of different levels of quality. For this purpose, this thesis study centers upon translation quality in order to comprehend how it impacts upon the variability of translation scores and upon the rating behaviors of raters. This study gains significance through its attempt to present generalizable and meaningful measurements that are considered to be relevant beyond individual contexts.

Purpose of the Study

When the studies carried out in the field of translation evaluation are examined, it can be seen that there has not been as much development as in the related studies conducted in fields such as Mathematics or Second Language Acquisition in the related studies (Angelelli & Jacobson, 2009). The fact that the number of empirical studies on the measurement of translation is inadequate in general displays the need for more detailed research on this specific subject matter (Vehmas-Lehto, 2008).

In experimental studies conducted previously, it was determined that the raters gave very different scores to the same translations (Speck, 2009). This situation shows that teachers mostly use subjective methods in evaluating translation and that the assessment is not sound, reliable or valid. The main issue in evaluating the quality of translation is to reduce the bias in the evaluation and to discover creative methods to increase the validity, reliability and consistency of the models. For this reason, recently, the primary aim of the translation quality assessment field has been to achieve inter-rater objectivity by minimizing subjectivity (Martínez-Mateo, Montero Martínez & Moya Guijarro, 2017).

Therefore, the purpose of this study was to investigate the impact of translation quality on raters' rating behaviors and on the variability of scores. Also, this study attempted to gain an insight into the different decision-making strategies and behaviors that raters adopt during the procedure of assessment. Following a research design of mixed-methods, the reliability and variability of the scores given to translations by the raters were investigated quantitatively by the application of G-theory framework. The raters' decision-making behaviors were examined through the use of qualitative data set obtained from thinking aloud protocols (TAPs).

The questions that guided this study were as follows:

1. Are there any significant differences among the holistic scores assigned to low- and high-quality translations?

2. What is the score variability between and among raters' scores assigned to each component and sub-component of the rubric?
3. What are the sources of score variation that contribute to the score variability of the holistic scores assigned to high- and low-quality translation papers?
4. Do the variables of gender and rating experience have an impact on the reliability of the scores assigned to high- and low-quality translation papers?
5. How do raters make decisions while assigning scores to different quality translations holistically?

Significance of the Study

Nowadays, translation used as a tool for foreign language teaching is gaining significance and it is considered to be an applicable method in teaching foreign languages to students (Cook, 2010). This situation increases the support for the teaching of translation in many different settings including universities (Ebbert-Hübner & Maas, 2017). In spite of the fact that the majority of studies conducted in the relevant field are theoretical, some experimental studies have examined the attitudes of instructors and learners towards the use of native language and translation in foreign language courses (Kelly & Bruen, 2015; Machida, 2008). The findings of these studies have shown that both teachers and students consider translation activities to be one of the facilitative language teaching tools and that they agree that the translation task helps learners use the target language properly.

According to Martinez Melis and Albir (2001), translation quality can be divided into three categories: 1) assessment of published literary and religious works, 2) assessment of professional translators' proficiency levels, 3) assessment of translations produced by students in educational contexts. It is the evaluation of student translations category that is the most neglected among these areas. Although a variety of criteria and models have been suggested for translation evaluation, almost none of these have been produced specifically to assess the translation products of undergraduate students (Medadian & Mahabadi, 2015). Given the impossibility of purely objective assessment in translation, subjectivity can be reduced as much as possible by the application of a transparent method and well-defined evaluation criteria (Martinez-Mateo, 2016).

While the main purpose of the evaluation in the context of education is to measure the skills acquired in the target language by examining students' translations, teachers have difficulty in applying specific models and evaluation criteria due to lack of time and resources when measuring a language skill. For this reason, teachers mostly prefer subjective evaluation methods based on their intuition when evaluating the translations of the students (Medadian &

Mahabadi, 2015). In the light of this, it is important to ensure that useful, valid and reliable measurement tools are made available to teachers for evaluating student translations that evaluation can be carried out objectively. It is essential to establish standard criteria that emphasize scientific objectivity and reliability in order to prevent unconfirmed evaluation judgments (Rabadan *et al.*, 2009).

Also, the absence of empirical research in the translation assessment area investigating variability of scoring and reliability of translation assessment is another inspiration for this research. Moreover, to the best knowledge of the author, in the field of translation assessment, there have been no attempts to examine the differences between different qualities (high-low) of student translations and the impact of translation quality on scoring reliability and variability of assessment.

Additionally, this study is considered to be important regarding its methodology. Although many quantitative studies apply classical test theory for data analysis, this study relied on G-theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972), which is a theoretical framework used for quantitative data analysis, in order to determine the variation of rater and scoring variability (Shavelson & Webb, 1991). Also, think aloud protocols (TAPs) were applied by the raters during the assessment procedure for some predetermined translation papers in order to obtain qualitative data that would give insights into the decision-making behaviors applied by the raters.

Therefore, this study aimed to examine variation of rater in translation assessment by putting specific emphasis on the impact of translation quality upon variability in rating. In addition, it attempted to provide insights into the sources that affect raters' decision-making processes, even when they used the same holistic rubric. The findings of the study will address its implications and suggestions for the standardization of translation assessment and for the regulation of institutional assessment procedures.

Limitations of the Study

In the present study, the lack of a thorough rater training procedure might have caused variations in scoring among raters. Although the necessary information was provided before the scoring process regarding the criteria included in the holistic rubric, it was observed that the rater training given was insufficient to obtain fairer ratings. Thus, providing more in-depth rater training regarding the application of the holistic rubric might have yielded more accurate and reliable scoring by reducing rater bias considerably (Barkaoui, 2010). Also, in this study, scoring time and conditions were not controlled by the researcher; raters were allowed to take

translation papers to their homes and they were given three months to complete the assessment task. However, performing the evaluation task under different conditions (e.g., physical/mental exhaustion, personal issues, emotional factors etc.) and in different time periods may have significant impacts on inter-rater reliability (Barkaoui, 2010).

Further, in this study, four different texts obtained from an online newspaper were given to students to translate. Previous research has revealed that the type of writing task might have an impact on score variability (Gebril, 2009; Şahan, 2018; Weigle, 1999). Therefore, the use of different text types might have caused students to perform differently. Also, the raters might have followed different rating procedures and exhibited different tendencies as a consequence of the use of different text types. Moreover, G-study analyses performed on the translation papers yielded high residual variances for high-quality (28.4%) and low-quality (34.7%) translation papers, showing that some unobserved errors and other hidden facets contributed to the score variability (Brennan, 2001; Şahan, 2018). Large contribution of residual to score variability is not a desired situation in G-studies, since it decreases the possibility of explaining observed sources of variance.

Another limitation of the current study concerns thinking aloud protocols, since two of the raters participating in the study failed to complete the thinking aloud tasks for various reasons. Although all the raters were provided with the necessary guidelines and a sample thinking-aloud video, 20% of the protocols could not be performed in the desired format. In addition, in spite of the desired degree of agreement between the two encoders (.82 $p < .001$), misinterpretation of the raters' utterances might have caused problematic coding (Cumming *et al.*, 2002). Also, 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 felt under pressure while thinking aloud, which might have caused problems related to the quality of the verbal protocols (Barkaoui, 2010). Moreover, raters' thinking aloud procedures might have been affected by the experimental nature of the current research, resulting in variability in scores (Şahan, 2018).

Definitions of Key Terms

The key terms which are regarded to be prominent in this research are as follows:

Translation assessment – it is defined as evaluating a translation by assigning scores to the product and interpreting on it. Various terms such as “grading,” “marking,” “rating” and “scoring” can be interchangeably used in order to refer to the procedure of assessment.

Rater – it refers to the evaluators assigning scores to written products.

Rater behavior – rater behavior can be defined as different procedures used by a rater to decide on the merit of a written product (translations in the context of the present study).

Holistic rubric – It refers to a rating scale focusing on the product or process as a whole and allowing for a very general evaluation without disrupting performance (Birel & Albuz, 2014)

Analytic rubric – It refers to a rating scale dividing the product or process into various dimensions, allowing each skill involved in performance to be evaluated independently (Birel & Albuz, 2014).

Object of measurement – it can be defined as the entity under investigation. In this study, translation papers obtained from EFL students are the measurement objects (Shavelson & Webb, 1991).

Facet – it refers to a specific aspect of a measurement method (Shavelson & Webb, 1991). Potential sources of measurement error are rater and translation quality in this research.

Condition – it can be defined as levels of a facet (e.g., for the quality facet: high-quality, medium-quality, etc.; Shavelson & Webb, 1991).

Universe – it refers to the overall number of conditions of a facet or amalgamation of facets (as in an interaction) (e.g., universe of raters, universe of items) (Shavelson & Webb, 1991).

Universe of generalization – it can be defined as the universe of conditions of a facet to which an evaluator wants to generalize (Shavelson & Webb, 1991).

Generalizability- (G-) study – it is a framework used for evaluating the relative significance of various measurement error sources and investigating the impact of changes in the measurement design (e.g., different qualities of essays; Brennan, 2001).

Decision- (D-) study – it deals with the practical application of a measurement procedure by integrating the ideal design to ensure the interpretation of the reliability of scores (Brennan, 2001).

Organization of Thesis

This thesis study is composed of five chapters: introduction, review of literature, methodology, results, and discussion and conclusion. Chapter one deals first with the issues of the use of translation as an assessment tool in departments of Foreign Language Education in Turkey and of how translation papers of students are evaluated by instructors. Then, the section addresses the relevant studies conducted on the validity and reliability of translation tests.

Chapter two presents a survey of existing approaches to translation assessment and discusses some major translation quality assessment models. Also, the chapter reviews the existing empirical research on translation quality assessment with a language oriented approach. Chapter three gives details about the design of the research and theoretical background of the study. Also, it describes the participants, instruments used for data collection and processes of data collection. Then, it gives information about the procedures of data preparation and analysis respectively. In chapter four, the results of the study are reported to answer the quantitative and qualitative research questions. The section ends with a summary of the findings. Chapter five presents a discussion of the findings, and emphasizes methodological and pedagogical implications.



CHAPTER TWO

Review of Literature

This chapter attempts to present a review of existing literature relevant to the use of translation tasks and tests in writing performance assessment. Firstly, the section briefly touches upon the issues of translation and translation quality, and it continues with an overview of assessment and subjectivity in translation quality assessment (TQA). Secondly, the issues of reliability, validity and fairness in translation quality assessment are discussed. Thirdly, the section reviews some approaches and major models in TQA. Lastly, empirical studies conducted on the impact of translation on second language (L2) acquisition and accuracy, comparison of translation tasks with other monolingual and bilingual tasks, and examinations of the validity and reliability of translation tests are reviewed, followed by a summary and identification of research gaps in the field of TQA.

Translation and Translation Quality

Various definitions of translation have been made previously by many linguists and scholars, but there is a shared definition upon which they all agree that translation is a practice of transfer, change, substitution, and reproduction. This common view can be traced back through the following definitions of the concept.

According to Catford, translation is a substitution process in which a text in a language is replaced with textual material (equivalent) in another language (1965, p.1). On the other hand, Newmark defines it as a procedure of conveying the intended message of a source text in one language to a text in another without changing the meaning (1988, p.5). For Newmark (1991), translation is a cover term including any transferring method from oral to written or from written to oral. It is the elegant and concise action of reproducing the message in the source language by using the most natural equivalent words in another language.

Munday (2001) also highlights the concept of change, claiming that translation is a process during which the translator changes an original written text in a source language into another written text in the target language. However, Basnett (1991) explains the concept more broadly and she emphasizes aspects of surface meaning and structures of language. According to her, translation is the process of transforming a source text in a language into a different language in order to provide an almost equal surface meaning in the two by preserving the

structures of the original language as much as possible. Adopting a totally different point of view, Faiq (2005) elaborates the concept of translation beyond all the definitions given above. He states that translation is more than the rendering of a source text into a target text. It is a transmission of culture related aspects of an original written text in a certain verbal language into another written text in a target verbal language.

Furthermore, House defines two different translation types: overt translation and covert translation (1977, p. 189). An overt translation addresses its receptors implicitly by enabling them to realize that they are reading a translation, not a second original text. In overt translation, the source text is expressed within a new speech event by making it possible for the members of a target culture to observe and judge the text from outside. On the other hand, a covert translation can be defined as a functional equivalence to the original source text. A covert translation operates as an original text in the culture of the target audience. Although this type of translation has no particular culture of target audience, it has a more deceptive and less complex (psycholinguistically) frame than the overt translation (House, 2014, p. 252). Regarding translation quality assessment, covert translations are more difficult to judge since they are not as straightforward as the overt ones (House, 2014, p. 255).

In spite of the fact that the notion of quality is all around us, suggesting a particular definition of it is not such an easy task because what we refer as quality in one context might not be the same in another context. Based upon their previous experiences and personal judgments, each person develops a distinctive notion of quality. Thus, proposing a single definition of this concept is rather difficult since it involves multiple aspects. As the Online Oxford Dictionary puts it, quality is “The standard of something as measured against other things of a similar kind; the degree of excellence of something”. On the other hand, in the Online Cambridge Dictionary, the term is defined as: “How good or bad something is”. And on the website of The American Society for Quality, it is defined as: “A subjective term for which each person has his or her own definition”.

Building upon the definitions of quality given above, it can be asserted that in order to decide on how good a translation is, each of the existing TQA models suggest some criteria that we need to follow. Considering the aforementioned definitions of translation, the existing TQA models suggest some criteria to be followed in order to assess the quality of a translation.

There has always been a close connection between translation and the notion of quality. Over the years, quality of translation has become a point of focus for scholars for a number of reasons that have been, for instance, aesthetic, political, religious, pedagogical, and economic (Williams, 2004, p.13). Globalization has led the world to become more dependent on effective

communication that is able to overcome cultural and linguistic barriers and promote mutual understanding. Regardless of the motives, individuals have always sought to produce what is regarded as a high quality translation.

Which characteristics can be ascribed to a high quality translation? How can we judge whether a translation is good or bad? Such questions have always been asked by many translation scholars over the decades. The available literature illuminates four main tendencies regarding translation quality. The first view, standing up for a subjective judgment, centers upon the criteria of equivalence between the source language and the target language (Cary & Jumpelt, 1963; Darbelnet, 1970; Nida, 1964). The second view can be characterized by its attempt to provide objective judgments. For example, Carroll (1966) used informativity measurements. Nida and Taber (1969) applied statistical analyses, cloze test techniques and tests including target language receivers. House (1981) and Williams (2009) created source text based quality assessment schemes. Leuven-Zwart (1990) performed evaluation systems to analyze literary translations. Following the emergence of the functionalist approach, scholars such as Nord (1992), Reiss (1983) and Vermeer (1978) put emphasis on the function of the text rather than the source language. These scholars highlighted the importance of the reported message in building communicative interaction between the translator and the target reader.

Unlike the older procedures and methods, a recent TQA trend not only deals with the function of a text and its reader/receiver, but it also takes into account the elements of efficiency, acceptability, and the hypothetical reader's needs (Gouadec, 2007; Mossop, 2001). Thus, it can be said that each different approach to translation yielded different notions of translation quality, and they all proposed various ways to measure it (House, 1997, p.1).

Assessment and Assessment Types

Colina defines assessment as 'a process by which information is collected relative to some known objective or goal' (2011, p.43). According to Burke, assessment refers to a process of collecting evidence or data for student learning to achieve instructional decisions (2010, p. 19). The notion of assessment functions as an umbrella term in the field of Translation Studies (Mcalaster, 2000, p. 231). Translation assessment, as the first component of the term determines the quality of a translated text (e.g. scoring translation papers either by letters or numerical grades). Analyzing a translated text to judge its value and appropriateness from a functional point of view constitutes the second component which is known as translation criticism. And the last component is translation analysis which concentrates upon comparison of a source text with its translated versions by considering equivalence of linguistic aspects (Mcalaster, 2000, p. 231). Mcalaster states that the terms assessment and evaluation can be used interchangeably

since they are synonymous (2000, p. 137). While Rodriguez associates the notion of evaluation with both translation criticism and translation quality assessment (2007, p. 5), Lambert asserts that criticism cannot be thought of apart from evaluation since it forms the basis for comparing a source text with a target text (1991, p. 31).

Kelly (2005) claims that assessment is carried out in two different ways in pedagogical settings: formative and summative assessment. While summative assessment is conducted at the end of a learning process in order to provide information for administrators and curriculum developers in taking decisions about adopting or funding a program, formative assessment takes place during the learning process to provide feedback regarding student or program improvement (Scriven, 1967). According to Bloom, Hastings and Madaus, summative assessment is performed to evaluate the efficiency of an educational plan, a curriculum or a program at the end of a learning process. On the other hand, formative assessment can be used for systematic evaluation of learning and teaching processes (1971, p. 117). The researchers also make a distinction between summative and formative tests. While formative tests are defined as achievement tests measuring certain units of learning, summative tests are used to measure a wide range of learning units (p. 137).

In the field of Translation Studies, two researchers (Kelly, 2005 and Colina, 2015) have explained the differences between summative and formative assessment types. Kelly defines summative assessment as “any assessment activity designed to award grades, permit progression to a higher level, or award a diploma” (2005, p.161). This assessment type, which aims at measuring learning outcomes of a course or program, is generally applied in the middle or at the end of a term. On the other hand, Colina (2015) argues that summative assessment is not commonly preferred in programs of translator training since it does not provide feedback on student progress. Kelly (2005) identifies formative assessment as an activity which improves learning by providing feedback on students’ progress. Focusing on the learning process, this assessment type is applied on multiple occasions from the beginning of a course to its end. Further, Colina (2015) states that formative assessment helps students learn from their mistakes, increase their performance and improve their current and future behaviors so as to achieve better results.

Stejskal mentions three factors in the translation assessment process; 'Translator', 'Process' and 'Product'. Two different methods are applied to evaluate the final product of the translated text: 1) the quantitative/bottom-up method based on the counting of existing errors in the translation 2) the qualitative/top-down method in which the translation product is considered as a whole (2006, p. 13). The quantitative assessment method focuses on the concept

of error by analyzing micro linguistic features at the sentence level and tries to determine the error according to a predetermined typology (Martinez-Mateo, 2016). In this analytical typology, performances under different headings are given different starting points and each translation error is subtracted from the total score of the relevant field. After the remaining points are collected, evaluation of the translation product is completed (Martinez-Mateo, 2016). The holistic method is centered upon the argument that the whole is greater than the sum of its parts (Johnson, Penny & Gordon, 2009). Raters are required to give a single score for a translation product and combine all essential features of translation to reach a general decision about the quality of translation (Al-Melhi, 2014). No matter which of these two methods is used, the criteria to be used in assessment should be clearly defined in advance (Stejskal, 2006).

Subjectivity and Prescriptive Judgment in Translation Assessment

Prescriptive judgment is a problem that is frequently faced while working on practical TQA (Lauscher, 2000). Prescriptive judgments are considered essential in translation practice contexts of translator training and quality control (Lauscher, 2000). However, House argues that the use of prescriptive judgment should be avoided since it requires an evaluator to know the beliefs and ideals of the persons involved in the translation process and since reconstruction of the decisions involved in composing a target text is a rather complex procedure. In her view, when an evaluator avoids use of prescriptive judgment, s/he shows respect to both the translation and the translator (1997, p.19). In contrast to this view, Lauscher (2000) claims that the complexity of the translation process can only be made visible through prescriptive judgment.

Weber (1988), a German sociologist, suggests that if a science is dealing with a social phenomenon, it should discriminate between value judgments and facts. According to Weber, value judgments and values can be examined by social sciences as social incidents; however, in a scientific context, it is not possible to derive value judgments from any scientific evidence (as cited in Lauscher, 2000, p.162).

One way of approaching the subject matter of prescriptive judgement could be the study of assessment process. In most of the TQA models, assessment is defined as a procedure in which a precise target text is compared to an ideal version of it (Lauscher, 2000). And there are three main factors affecting the translation evaluation process: the object, the evaluating subject and the ideal target text.

Regarding the subjects as evaluators, Pöchhacker suggests that they might have an impact on the way a target text is judged (e.g. translators, researchers, teachers, clients, etc.).

He states that the evaluator's background translation knowledge and experience are also factors that have an impact on how a target text is judged (1994, p.123). Reiss states that the object of evaluation is related to how a person interprets a target text. Since interpreting a text is an individual action, each evaluator's interpretation composes the evaluation object and it necessarily affects his or her judgment (1971, p.106). Also, the comparison of the sample target text with its ideal version affects the judgment. Each ideal version of a target text is created based on a general translating concept (House, 1997, p. 1). On the other hand, there is no single translating concept upon which any scholars agree. A review of translation history (e.g., Levefere, 1992; Venuti, 1995) reveals that translating related concepts showed difference over time. And these changes are encountered not only in intercultural, but also in within cultural dimensions. Lastly, a number of scholars (Ammann, 1990; Höning, 1998) have claimed that the context in which assessment is actualized has an impact on the judgment of a sample target text. The purpose of the judgement might show difference regarding the context (Ammann, 1990, p. 211). The assessment might be carried out for the purpose of determining a translator's qualifications, for providing feedback to students about their progress or for informing the reader in respect to the translation quality of a work of fiction, etc. (Höning, 1998).

Since translation can be identified as a creative procedure rather than a mathematical one based on formulas, a source text might have a number of acceptable translated versions in a particular target language. Although each of these versions can be ranked, ranking might differ from one evaluator to another: therefore, perfect objectivity cannot be achieved (Mohacsi-Gorove, 2014). As translation quality is, inherently, an ambiguous concept, opinions demanding objective measurements have been verbalized recently. According to Angelelli (2009, p.1), an objective measurement of translation quality is essential in an academic setting in educational programs for formative evaluation, such as 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. Thus, in educational contexts, it is essential to examine the procedure of TQA further in order to identify the factors that impact upon evaluators' decision-making. Although it is a well-known fact that assessment of translation quality can never be ultimately objective, Kis (2008) argues that a translation can be regarded as acceptable when more than one reviewer assumes that it is so. This thesis study partly aims at testing this argument by giving insights into the complex reality of TQA.

Since fair judgment has a crucial role in obtaining generalizable and meaningful measurements, the following section reviews the issues of validity, reliability and fairness in the assessment of L2 writing tasks.

Reliability, Validity and Fairness Issues in Translation Assessment

Scoring procedures provide a basis for distinguishing between subjective and objective tests (Bachman, 1990). While the answers of a test taker in an objective test are scored according to a pre-prepared answer key, in a subjective test the correctness of the answers is determined according to the interpretation of the evaluators (Bachman, 1990). In subjective tests, evaluators' subjective interpretations are determinative in assessing the performance of learners, which raises the question of how reliable such evaluations can be. Since writing assignments are assumed to be subjective by nature, this standpoint seems particularly acceptable. That is to say, when a score is assigned to a writing task, it is not regarded as the result of the interactive relationship between the test and the test-taker, but as the outcome of a number of factors that may have introduced errors of measurement into the writing assessment, such as the task, the test-taker, the evaluators, and the scoring rubric (Kenyon, 1992; Mcnamara, 1996). Thus, reliability is related to the extent to which a test score (a student's performance) is impacted upon by the measuring error, rather than to the linguistic ability to be measured (Bachman, 1990). In other words, the concept of reliability seeks for consistent scores among different evaluators, times, contexts, test types and other measurement characteristics (Weigle, 2002).

Weigle (2002) mentions two types of inconsistency in measurement: inter-rater reliability and intra-rater reliability. While inter-rater reliability is related to the consistency of the scores given to the same task by two or more evaluators, intra-rater reliability can be defined as the consistency of the scores that a single rater assigns to repeated administrations of a task. Although ensuring inter-rater reliability in writing assessment is a difficult task, White (1994) suggests several practices to improve it: a) the writing tasks should be graded in a controlled reading, namely the scoring circumstances should be same for all the evaluators, b) the adherence of each of the raters to the pre-determined rating criteria should be monitored, c) grading sessions should be evaluated and discrimination should be made between a reliable and an unreliable evaluator.

According to the principles and practices of measurement theory, the reliability of a test is a prerequisite for its validity; that is, if a test is not reliable, valid results cannot be drawn from scoring it (Anckaert, Eyckmans, Justens, & Segers, 2013, p. 80). The American Psychology Association defines validity as the “relevance and usefulness of certain inferences derived from test scores” (Bachman, 1990, p. 243). For a translation test to be valid, it must measure the correct structure, e.g., translation ability (Angelelli, 2009, p. 24). While reliability is related to the test scores' quality, validity is concerned with the interpretation quality of a test

(Bachman, 1990). Bachman (1990) considers reliability and validity as complementary structures, because both of the structures aim to minimize the impact of measuring error and maximize the impact of the language skills we intend to measure.

Although fairness is a desired characteristic in writing assessment, it is difficult to achieve because of several factors affecting the reliability of scores (Breland, 1983). When individual and externally guided value judgments are considered, we see that translation evaluation is a problematic procedure (House, 2001). In fact, it is very difficult to achieve a final judgment that meets the criteria of scientific objectivity when measuring translation quality. Although subjective judgment is inevitable in translation evaluation since the procedure is carried out by a person, this situation necessarily points to a need for objectivity (House, 2001). It is not likely that an absolute reliability can be achieved in writing assessment (Hughes, 2003). However, taking some precautions, such as double-evaluating the translations, training the raters in how to grade the papers and using rubrics for scoring, might enhance the reliability of assessment (Bachman & Palmer, 2010). In this regard, the use of rubrics, in particular, in the assessment of written performance, such as translation allows measurement to be more consistent and objective (Sezer, 2005).

Approaches to Translation Quality Assessment

Subjective (anecdotal) and mentalist approaches to evaluating translations.

The association between the target text and the source text is what is essential for the main concept of translation quality. According to this approach, a target text can be considered successful, if only it transfers the originality of the source text by presenting an accurate and faithful equivalence for it (Newmark, 1991, p. 111). This accuracy should be sustained both on qualitative (e.g., the meaning should be given without any alterations), and quantitative (omission should be avoided, linguistic features of the original text should be represented) levels (Newmark, 1991, p. 111).

Regarding TQA, intuitive and subjective judgments form the basis for this approach. An individual who adopts this approach counts on his/her interpretation abilities, field knowledge and artistic-literary intuition when taking decisions about the quality of a translation (House, 2015, p. 9). Mentalist views are anecdotal judgments which are based on an individual's interpretation of how good or how bad a translation is. These judgments might result from simple emotions and impressions. Some examples of these judgments are: "Translation does not reflect the spirit of the original work", 'The soul of the original work is lost in this work', 'This translation is better than the original' (House, 2015, p. 9). It might be

thought that such statements made long ago on translation evaluation have become obsolete. However, scholars of the “neohermeneutic” school who justifies the validity of ambiguous and subjective interpretations still accept and use this approach in translation evaluation (e.g., Prunc, 2007; Stolz, 2011).

In brief, most of the approaches adopting anecdotal translation assessment highlight the idea that translation quality depends mainly on the translator’s cultural and linguistic knowledge and on his subjective and intuitive interpretation and decision-making (Cuellar, 2002). On the other hand, House claims that the subjective and intuitive interpretations of the mentalist views are not appropriate for the assessment task in which it is argued why, how and when a translation is assumed to be good (2001, p. 244).

Behavioral approaches to judging translation quality.

Contrary to the subjective-mentalist views that handle the TQA concept intuitively, the behavioral view describes the mental actions of translators as an unknown black box and it advocates that translation assessment should be carried out in a more scientific way (House, 2001). This approach influenced by American structuralism and behaviorism, is often regarded as associated with Nida's (1964) pioneering work. According to Nida (1964), the response or reaction of the reader to a translated text should be taken as the main criterion in TQA, and he argued that translation assessment should be actualized in accordance with his proposed principle of 'dynamic equivalence of translation'. Nida claims that the reaction displayed by the readers to a translation should be equivalent to the reaction of the source text readers. In this case, according to the behavioral approach, a good translation should have equal “intelligibility”, “informativeness” and “reader response”. However, House (2001) argues that measuring such complex structures is not possible because it is very difficult to consider these structures as criteria in assessment of any translation. Also, she states that although a translation is assumed to provide equivalence to the original text, it still raises the questions: to what degree is equivalence of response met? And is it possible to empirically test whether all requirements are met? Furthermore, House claims that the source text is largely ignored in behavioral approaches. In other words, the relationship between the original text and the translated text is not taken into consideration (2015, p. 11). For House, the main weakness of behavioral approaches to judging translation quality is that they all ignore the human mind, ‘such that, for instance, tests involving expert judges must take criteria for granted that need to be developed and made explicit in the first place’ (1997, p. 5).

Functionalist approaches to assess translations.

Micro-textual linguistic models centering upon the characteristics of a source text and their transference to a target text gave place to functionalist approaches that primarily focus on macro-textual features. A functional approach, by taking into account the complexity and function of a translation, reframes the text in order to provide a precise rendition of the cultural background, time, medium and place of the agents (Nord, 2007). These situations are described as 'situation in culture' (Nord, 2007). Within this approach, how to translate the norms of a target culture in a translation is a substantial criterion in TQA (Nida, 1982, p. 64). For this approach, it is not essential to translate the linguistic structures or words in an original text, but to reflect each individual communicative expression in the speech event.

The pioneers of this approach (e.g., Reiss & Vermeer, 1984) suggest that the most essential factor in the assessment of translation quality is the 'purpose of translation'. One of the major theories developed under the functionalist approach, Skopos theory, introduced by Vermeer (1978), draws upon the action theory by highlighting the significance of identifying the main objective of translation. In this regard, the source text is dealt with as the raw material of the translator through which the main purpose is recoded into the target text. Since this purpose shows variety, Vermeer argues that there is no single accurate translation acceptable for the original text.

Although the notion of "function" is regarded as essential, House (2001) criticizes this approach for not providing a satisfactory explanation of it. This notion might be depicted as the real-world impact of translation (House, 2001). However, this approach does not fairly explain how to determine the notional equivalence or adequacy of a translation. What is ignored here is the fact that a translation cannot be thought independent from the original text from which it is produced. By its nature, a translation is synchronically based upon the purpose and the language-culture context of the source text. Translation quality assessment cannot be considered apart from the cultural values that a source text bears; this is far beyond describing defects and strong points in translations (Yamini & Abdi, 2009). Thus, the functional approach does not seem to provide adequate criteria for the assessment of translation products (House, 2001).

Major Translation Quality Assessment Models

Since translation models are based on particular translation theories, assessment of a translated work's quality should be carried out based on a definable, practicable, and testable model (Williams, 2004). Numerous attempts have been made to construct a model for the

assessment of the quality of translation works (House, 1988, Williams, 2004). Each model of translation defines a distinct aspect against the other. For example, while grammatical model deals with the linguistic patterns of a translated text, the cultural model focuses on the communicative aspect of the same text.

According to Williams (2004), TQA models can be divided into two main categories: 1) there are models with a quantitative measure and 2) there are non-quantitative, text-logical models. House (1998), mentions three categories of TQA model: the first category includes pre-linguistic models, in which subjective judgments and ambiguous expressions related to the quality of a translated text play a role. The second category comprises psycholinguistic models, in which the translation quality of a translated work is assessed with respect to the impact of a translation on its readers. And the third one involves source-text based models that aim to constitute some linguistic measures in order to refer both to the source text and the target text. The following section discusses some of the major TQA models.

Translation quality assessment model of Larose.

Robert Larose argues that objective of a translation is central for its quality assessment. The focus of this model is the assessment of effectiveness; that is, the extent to which the purpose of the translator corresponds to the intentions of the author of the original text (Drugan, 2013). According to Larose, each translation should be assessed in terms of its relevance to the purpose of the original author, rather than the intentions of the evaluator (Williams, 2004, p. 9). After the objectives are determined, the evaluator identifies the assessment criteria and limitations such as cost, time, social values and requirements of clients.

Larose's approach to TQA includes both textual and extra-textual elements. Unlike the TQA model proposed by House, it deals with the impact of the experience and gender of the evaluator and of the conditions in which the evaluation is actualized (Drugan, 2013). Larose presents his assessment model in three basic levels.

1. The lowest level, which deals with expression forms and lexical and syntactic elements.
2. Macro-structural level, which concentrates upon evaluation of discourse, content and cohesion features.
3. Super-structural level, which refers to the evaluation of the overall structure of the translation's narration and argumentation.

In this approach, these three levels show the seriousness of the errors in a translation. For example, an error made in the third level is more serious than an error made in the first level

(Drugan, 2013, p.55). Application of the model starts with identification of extra textual elements related to the purpose of translation by comparing the target text with the original one. Then, these two texts are separately evaluated at all three levels in terms of the purpose of the original text and what the translator has in his mind (Williams, 2004).

Since translation purposes differ from one text to another and there are various ways of translating a text, this approach does not seem applicable to the evaluation of all translation texts (Drugan, 2013). Although, Larose's TQA model centers upon the significance of extra textual factors, he does not provide enough details about them (Drugan, 2013).

Williams' argumentation-centered assessment model.

The theory of argumentation was first introduced by Toulmin in order to provide a practical explanation for the continuum of revealing an argumentation and justifying it through a number of constituents (Kharmandar, 2017). Drawing upon the existing argumentation theory, Williams proposes a new framework to assess translation quality. Williams asserts that present TQA models mostly deal with micro-textual evaluation criteria such as morpho-syntactic and discrete lexical features and he believes that application of these units is problematic when it comes to the evaluation of long passages (2004, p. 17). Integrating macro-textual elements and pragmatic aspects into his assessment model, Williams attempts to minimize the subjectiveness and randomness of the evaluators' decisions. In this approach, the main focus is on determining whether the argument of the source text can successfully be transferred to the target text. In this model, the assessment is carried out in four stages:

1. In the first stage, the original text is analyzed in order to frame a schema for its main argument, then the essential messages of the text are identified.
2. In the second stage, the translated text is analyzed apart from the source text for the assessment of its overall coherence. Also, probable problems related to its acceptability and readability are determined.
3. In the third stage, the translated text is compared with the source text and following parameters are evaluated:
 - a) Schema of argument
 - b) Argument types
 - c) Relations of organization
 - d) Arrangement
 - e) Speech figures
 - f) Indicators of inference
 - g) Narrative strategies

4. In the last stage of the assessment model, a declaration of the overall quality of the translation is made (Williams, 2004, p. 73).

According to Williams (2009), Toulmin's translation model has two essential elements, 'claim' and 'ground'. While claim which is given in the purpose constitutes the most important part of argument, grounds support the claim by providing some rational information such as proof, observation or facts. However, unlike Toulmin's model, Williams makes a distinction between critical, major and minor errors that are defined with respect to the usability of a translation (2004, p.133). Williams asserts that his model's final version serves as an approach that includes all the elements essential for instrumental translation, and what is central in this model is the function and the use of a translation (2004, p.152).

In order to test the applicability and efficiency of the model, Williams preferred professionally translated passages as samples. However, just a few particular text types were used in the testing phase of the model, and only one language pair (from English to French) was evaluated. Further, Kharmandar (2017) argues that Williams' TQA model has two basic problems in terms of its application. Firstly, it does not shed light on how to analyze the complex structures of arguments. And secondly, although it is a model proposed for instrumental translation, it mostly deals with literary translation. Drugan also questions the validity of this model since it is not possible for all text types to have equal argument structures (2013, p. 63)

Al-Qinai's translation quality assessment model.

Al-Qinai (2000) suggests that, when evaluating translated texts, some measurements should be taken into consideration such as the extent to which the target text meets the syntactic, semantic and pragmatic functions of the source text, including cultural aspects and the capacities for expression of both the target and the source language. According to him, translation functions in utterances rather than sentences. And, even though it invalidates semantic meaning, a pragmatic level equality is what is looked for. Al-Qinai proposed an eclectic assessment model that can be tested empirically in the analysis of the language features of the source and target texts, in both pre- and post-translational quality measurement phases of the target text. In order to provide an objective explanation regarding the relative equivalence of target text and source text, Al-Qinai tested his model on a particular advertisement text.

Al-Qinai (2000) describes translation as a complicated hermeneutic procedure in which the intentions of the source text author are interpreted intuitively. Nonetheless, he suggests a

number of criteria proposed by other scholars (House, 1997; Newmark, 1988; Steiner, 1994) for an objective translation quality measurement. These are:

- a) Tenor and typology of text: comparing the source text with the target text in terms of textual functions and their narrative and linguistic structures.
- b) Correspondence: evaluating the target text regarding the aspects of arrangement, paragraph division, expression of the headings...etc.
- c) Cohesion: looking for the suitable use of conjunctions, deixis, ellipses and substitution.
- d) Pragmatic equivalence: examining whether the target text reflects the same intended impact of the source text.
- e) Register: comparing the target text with the source text regarding appropriate use of idioms, jargon, loanwords, phrases, connotations, collocations.
- f) Equivalence of grammatical aspects: order of words, structures of sentences, agreement of tenses and aspect (Al-Qinai, 2000, p. 499).

Al-Qinai (2000) does not provide any justification for his preference for these parameters or for the exclusion of the others suggested by the theorists. For the practical application of his model, Al-Qinai defines each of the above elements by providing information about their theoretical background and then compares the target text with the source text in terms of areas of correspondence and divergence. After each of the elements has been applied in detail to the quality assessment of the translation, he evaluates the target text with a “holistic view” and provides judgments about its level of the acceptability. Al-Qinai's translation quality assessment model is criticized for not yielding rational explanations for the notion of overall adequacy and for not providing criteria regarding the selection of assessment parameters (Drugan, 2013, p. 59).

House's translation quality assessment model.

In the construction of her first translation quality assessment model, House (1977) drew upon a number of theories such as functional and systemic theories, pragmatic language theories, and equivalence and register theories. House's early model, which is based upon the notion of equivalence, centers on the preservation of meaning or message across two different languages. Here, she highlights semantic, pragmatic and textual aspects of meaning which are considered to be essential for translation.

While semantic aspect of meaning refers to the association between words and the ‘things’ they refer to in a possible world which is constructed in the human mind, the pragmatic

aspect of meaning is related to the communicative use of utterances or sentences in any social context. Pragmatic meaning, in other words the illocutionary force of an utterance which is clarified by the context in which it is used, can be anticipated from grammatical aspects, such as intonation, stress, the order of words and performative verbs. House (1977) claims that although translation mainly deals with the way in which a language is used in particular occasions, evaluators mostly prefer to center upon the semantic aspects, disregarding the pragmatic meaning. House suggests that in certain situations it is more important to achieve pragmatic rather than semantic equivalence across the texts (1977, p. 35). Lastly, textual meaning is related to the formation of a cohesive translated text which constitutes equivalence to the original text in respect of ellipsis, substitutions, co-textual references, anaphora and theme sequences. Thus, the translated text should connect sentences cohesively by conveying the semantic meaning of the original text, and this is another aspect that evaluators neglect in translation assessment.

According to House (1977), after a detailed analysis of the original text to determine its function, what is sought should be the equivalence of function between the source text and the translated text. For House, each individual text is built in a specific situation, and clarification of a text's function requires us to refer to the situation itself by breaking down the notion of situation into particular dimensions:

- a) Field: Subject matter or social action.
- b) Tenor: Writer' or translator' stance, social role relationship and social attitude.
- c) Mode: Medium, participation and function.

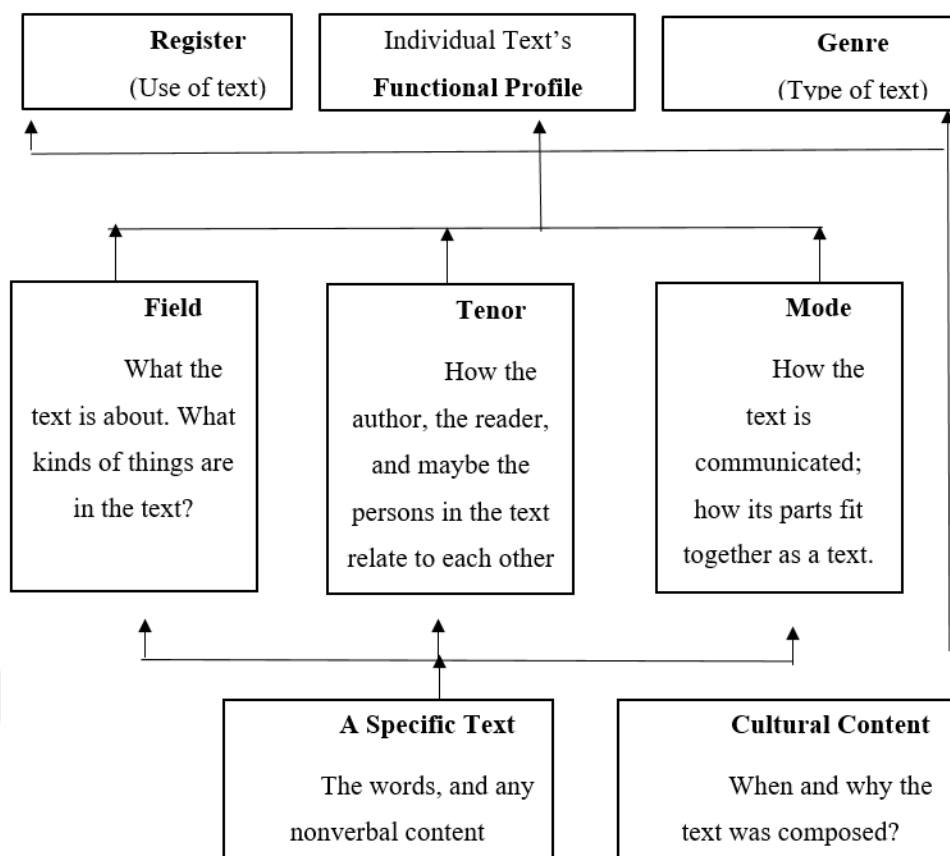


Figure 1. House's scheme for analyzing and comparing texts (1997, p.108).

Regarding error analysis, House claims that the different profiles of the original text and translated text should be prepared by the researchers when the profiles of these two texts do not match, and she mentions two categories of errors: overt and covert errors. While overt errors occur as a result of a non-dimensional mismatch, covert errors occur due to a situational dimension mismatch.

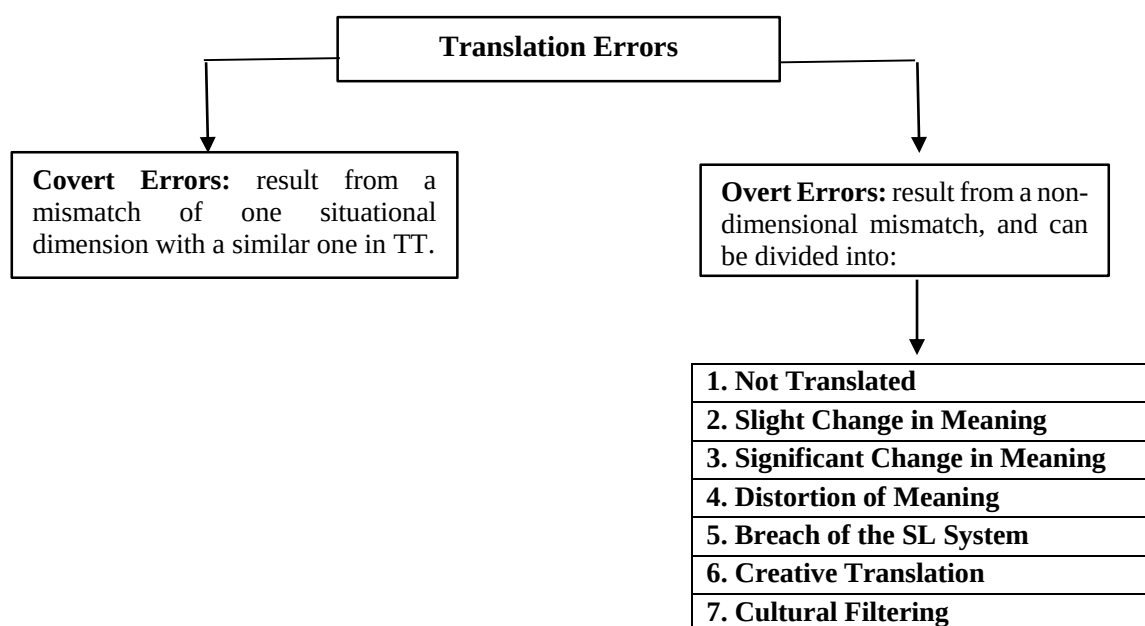


Figure 2. House's scheme for translation errors (1997, p.108)

Empirical Studies on Translation Quality Assessment

Researchers and scholars have paid great attention to translation research since the day it has emerged as a discipline (Khabir & Fumani, 2013). TQA as a subfield in translation studies is growing fast (Williams, 2004). What it deals with is the inter-relationships between the text translated from source text and the one translated into target text. The inter-relationships mentioned here include the linguistic characteristics of source and target texts such as lexis, syntax, grammar, semantics (Khabir & Fumani, 2013).

Theorists and critics in the field of translation studies have always been in agreement that any translated work should be assessed in terms of its quality considering its linguistic features. The term TQA essentially relies on the concept of evaluation. Scriven (1993) defines evaluation as identification of worth and merit. Thus, the primary reason for TQA to evaluate a work of translation is to decide on whether it is “good” or “weak” (Khabir & Fumani, 2013). Thus, it can be said that “the evaluation of translation as good or poor is the main concern of TQA studies” (Yamini & Abdi, 2009, p.2). Empirical studies conducted in the area of TQA have mostly centered upon the comparison of a source text and a target text in terms of their similarities and differences, and dealt with the use of translation as an assessment tool. Some of them have focused on the impact of translation on L2 learners’ accuracy.

Table 1. *Summary of Reviewed Articles*

Focus of the study	Number of studies	Studies reviewed
Comparison of translation tasks with other monolingual and bilingual task types	12	Cohen & Brook-Carson (2001); Fumani (2013); Jamshidian & Mohammadi (2012); Kobayashi & Rinnert (1992); Lo (2016); Mohaghegh (2011); Rabadan, Labrador & Ramon (2009); Scherer & Wertheimer (1964); Sireci, Yang, Harter & Ehrlich (2006); Sharkas (2009); Vahiddatsjerdi, Hakimshafaai & Jannesaari (2008); Yamini & Abdi (2009)
Examining validity and reliability of translation tests	8	Colina (2008); El-Banna (1993); Eyckman, Anckaert & Segers (2009); Ghonsooly (1993); Ito (2004); Neves (2002); Orozco (2000); Tavakoli, Shafiei & Hatam (2012)
Impact of translation on learners L2 acquisition	5	Calis & Dikilitas (2012); Kallkwist (1988); Laufer & Girsai (2008); Lee (2013); Prince (1996)
Impact of translation on L2 learners’ accuracy	7	Berggren (1972); Ghaiyoomian & Zarei (2015); Kallkwist (2008); Schjoldager (2003); Soleimani & Heidarikia (2017); Uzawa (1996); Vaezi & Mirzaei (2007)
Total	32	

Comparative studies on translation assessment.

Centering upon the comparison between translated writing and direct writing, most of the empirical studies in the literature have examined the use of translation tasks as compared to other monolingual task types (e.g., Cohen & Brook-Carson, 2001; Fumani, 2013; Jamshidian & Mohammadi, 2012; Kobayashi & Rinnert, 1992; Lo, 2016; Mohaghegh, 2011; Nobar & Navidpour, 2010; Rabadan, Labrador and Ramon, 2009; Scherer & Wertheimer, 1964; Sharkas, 2009; Sireci, Yang, Harter & Ehrlich, 2006; VahidDatsjerdi, Hakimshafaai & Jannesaari, 2008; Yamini & Abdi, 2009).

For example, in a study conducted by Kobayashi and Rinnert (1992) the translated writing and direct writing of advanced learners were compared in terms of their quality. The participants in the study were 48 fourth year university students, majoring in various fields at a state university in Japan. The proficiency levels of the students were determined as high or low through a standardized grammar test and an oral interview. These students were randomly assigned into two groups. The tasks were carried out during the regular class hours of the participants. First, one of the groups was required to write essays on a particular topic in their mother tongue (Japanese) and then to translate them into their L2 (English). In the meantime, the other group was asked to write essays on the given topic directly in L2. The following day, the groups were given reversed tasks and they were asked to write essays on a different topic. Three main aspects: 'organization', 'content' and 'style' were considered in measuring the quality of the papers. Each of the aspects was rated on a 5-point scale. The results of the research indicated that essays written in the translation mode were of a higher quality than the ones written directly in L2 especially on the syntactic complexity dimension. Also, it was found that students at low proficiency levels facilitated more from the translation task than the ones at higher levels did. As for the advantages of translation, the students reported that it helped them develop ideas more easily and express their thoughts and opinions more clearly.

In a simulated assessment context, Cohen and Brook-Carson (2001) examined the merit of translation used as an assessment tool in predicting L2 learners' writing ability. In accordance with this purpose, the researchers compared the direct and translated writings of students from the University of Miami. In total 39 intermediate level learners of French (9 male and 30 female) took part in the study. While 25 of the students were native speakers of English, 10 were bilingual Spanish-English and four had other mother tongues. Following a background information questionnaire, two writing prompts were administered to all of the participants. The prompts were obtained from a list including sample composition topics and in order to verify the translations' accuracy, native French and Spanish speakers translated them into their

first language (L1) and then back to English. The writing prompts were administered to the students in two sessions. The allocated time for each session was 30 minutes. In the first session, the participants were asked to write essays in their L1 and then to translate them into French. In the second session, all the students were required to write essays directly in French. Two experienced raters assigned points to the essays by using The Multitrait Scales. The essays of the students were evaluated regarding four aspects: expressions, transitions, clauses and grammar. The findings of the study indicated that two-thirds of the participants performed better in direct writing than translated writing. However, no significant difference was found between the two writing tasks regarding the learners' L2 grammatical accuracy. Also, the participating students reported that under time constraints writing directly in L2 was more effective. The researchers suggested that the use of different tools in the assessment of writing might be helpful in avoiding biased results.

An early experimental study was carried out by Scherer and Wertheimer (1964) in order to compare the relative merits of two second language teaching methods, The Grammar Translation Method and the Audio-lingual Method. A total of 300 freshmen students majoring in German at the University of Colorado were included in the study. While half of the participants received instruction using Audiolingualism, the other half was instructed through the Grammar Translation Method. The experiment lasted for two years and the researchers monitored the distinctions between the audio-lingual and cognitive code methods. During the process, the students were administered large batteries of tests and various dimensions of the methods were compared and contrasted. Although the researchers expected the Audio-lingual Method to be superior in all the dimensions, the results of the study revealed no significant difference between the groups in relation to L2 translation and reading skills. Students who were taught by the Grammar Translation Method were found to perform better than those taught by the Audio-lingualism in L1 translation tasks (from German into English). However, the Audio-lingual group outperformed the traditional group in speaking tasks.

In a longitudinal study focusing on the outcomes of translation tasks, Lo (2016) compared L2 learners' reactions to a monolingual writing task and particular translation tasks. For this purpose, 26 third-year college students (17 female and 9 male) attending a public administration program at the Institute of Macao Polytechnic were involved in the study. The participants, whose ages ranged from 20 to 22, were taking an authentic English course at the time of the study and they all had similar language backgrounds (L1 Chinese and L2 English). The researcher randomly divided them into two groups. A set of instruments including writing and translation drafts, questionnaires and class-discussion transcripts were used for data

collection and the procedure lasted for two semesters (2014-2015). The research was carried out in two experimental phases. In the first phase, half of the participants were given translation tasks and the other half worked on monolingual writing tasks. In the second phase of the study, the groups were exposed to reversed types of task. Thus, by swapping roles both groups worked on two different task types. The results of the comparison analysis indicated that the students who were exposed to monolingual writing tasks made more lexical and grammatical errors than the ones who worked on translation tasks. Also, it was determined that the students who experienced translation tasks developed more positive attitudes towards use of translation in foreign language (FL) classrooms.

Using the methodology of differential item functioning, Sireci *et al.* (2006) evaluated survey items which were translated into various languages in order to prove the hypothesis that developments in the process of test adaptation yield better quality translations. The data of the study were obtained from a survey (including 13 items) used by an international firm for employee engagement. The instrument was originally developed in English and it was then translated into 44 different languages. In the study, translated versions of the survey in three different languages (U.S. English, Italian and Polish) were examined. The researchers compared two different time points (1999 and 2003) at which the survey was administered in order to identify improvements occurring in the process of translation. For the initial translation examination, 2500 English, 531 Italian and 2465 Polish survey samples were analyzed. And for the revised translation analysis, 2500 English 510 Italian and 1739 Polish survey samples were evaluated. The following steps were taken in the process of expanded translation: (1) initial translations were administered, (2) translations were reviewed and a discussion was held between research firm personnel and the translators, (3) a number of independent translators carried out back-translation process, (4) original and back translations were reviewed, (5) final verification was conducted. For each of the item that the survey included, separate logistic regression and ANCOVA analyses were performed. The researchers found that back-translation produced results that were more comparable to the original English measure for many of the survey items. In other words, findings indicated that similarity of the translated items to their counterparts of English was improved through a more rigorous translation.

In a study focusing on the translation of poetry, VahidDatsjerdi, Hakimshafaai and Jannesaari (2008) tried to construct an applicable model of translation assessment that could be applied to measure the quality of translation in poetic genre. In this product- oriented study, the researchers contrasted a piece of Persian poetry with its translated English version. The two texts were compared at both textual and extra textual levels. At the textual level, similarities

and differences (in terms of linguistic aspects) between the target text and the source text were determined. At the extra-textual level, the two versions were handled according to their cultural framework. The findings of the study were analyzed descriptively. The results revealed that in the assessment of poetry translations, some unresolved issues and difficulties are encountered in relation to lack of authenticity and to aesthetic aspects of the original text in the translated version. At the end of the study, the researchers proposed a model for this problematic act.

Implementing the model proposed by VahidDatsjerdi *et al.* (2008), Jamshidian and Mohammadi (2012) conducted a study to determine the closeness between a piece of Persian poetry and its translated English version, regarding their poetical and grammatical features. In the study, the researchers analyzed the similarities and differences between the source text and the target text descriptively at an extra-textual level. Having compared the rendering with the source text, the researchers concluded that the target text was grammatically close to the original text. Further, the target text was successful in conveying the message of the source text, both literally and pragmatically. Thus, it was concluded that a translator's expertise, and cultural and background knowledge are key factors in the success of a translated work. In another study conducted by Nobar and Navidpoor (2010), the same model proposed by Vahiddastjerdi *et al.* (2008), was used in order to compare a piece of poetry by Rumi's Ghazals to its translated version by Shahriai (1998). These two texts were examined at both textual and extratextual levels in order to determine the closeness between them. The researchers found that while textual elements of the target text showed closeness to the original text, it was unsuccessful in preserving the cultural aspects of the source text properly. Also, it was found that exclusion of the cultural aspects of the source text spoilt the coherence of the target text, leading to the reader experiencing a difficulty in understanding the message.

In another study, Sharkas (2009) tried to determine potential problems that translators of science articles encounter and examined whether they could produce adequate translations. Adopting a translation analysis method proposed by Hervey and Higgins (1992), the researcher focused on translation loss in order to measure the quality of a target text in terms of different textual levels such as 'genre, cultural, semantic, formal and varietal' (p.48). Popular English science articles and their Arabic versions were compared to each other regarding these five features. The translation brief and genre characteristics were also taken into account during the process of assessment. It was found that translators mostly facilitated from imagery and analogies to characterize genre and that in giving complicated information they followed a process from known to unknown. However, the researcher identified translation loss on the semantic and formal levels. In a similar study, Mohaghegh (2011) followed the text-based

approach to translation assessment introduced by Reiss (2000) to evaluate an English translation by Clarke (1891) of Ghazal No.167 of Hafez, according to its linguistic elements such as semantic, lexical, grammatical and stylistic features. In a comparison of the original poem with its translated version, it was found that at the linguistic level there arose some translation problems resulting from miss-reading of the source text by the translator. Thus, the researcher determined that some of the idioms in the source text were translated unintelligibly because of a literal rendering.

In another comparative study, Yamini and Abdi (2009) examined the quality of a Persian translation of William Shakespeare's *Macbeth* by applying House's (1998) TQA model. In the study, some samples of the original text and the translated version were chosen randomly and then these texts were analyzed in detail. The errors (both major and minor) were determined and categorized by the researchers. Some statistical procedures were performed in order to determine the frequency of error occurrences. Then, the Chi-Square procedure was performed to determine significant differences between the observed and the expected (overtly erroneous and covertly erroneous errors) frequency of the errors. The results of the analysis revealed significant differences between these two types of errors. As a result, this specific piece of translation was not found to be compatible with the hypothesis of House (1998) that a literary work should be translated overtly and deviations from the source text should be accepted as errors.

Fumani (2013) examined Persian translated versions of English lyrics in terms of their quality, based on Hurtado's (1995) error translation model. The samples of lyrics were selected from six different music styles by the purposive sampling method. First, the researchers compared the Persian translations with the corresponding original lyrics in order to understand whether any erroneous renderings had been made during the translation process. The errors determined in the translated versions were divided into three categories: 'translation mistakes', 'translation major errors' and 'translation minor errors'. To assess the quality of the translations, two raters followed a scoring system. For each of the categories, the raters gave negative points of 2, 1 and 0.5. As the lyrics were graded out of 100, the sum of the negative points was subtracted from that total. When the raters gave the same score to the same translated works, they were recorded. And, when the ratings were different, a compromise was agreed between the raters. The mean scores for the music styles were calculated and a Chi-Square test was performed in order to find possible differences among the mean scores of the music styles. In the comparison of the translated lyrics' qualities with the original lyrics', no significant differences were determined.

In their study, Rabadan *et al.* (2009) presented a theoretical background to the notions of translation universals (simplification, source language interference and unique grammatical feature) and 'translationese' which can be described as a translation language produced under the influence of the original language. Applying a contrastive methodology, the researchers examined each of these three features of the translated language (Spanish) in three distinct case studies. Results for the source language (English) obtained from a corpus were compared to the results of these case studies in order to measure the language accuracy of the translated texts. Two monolingual corpora (English and Spanish) and one bilingual corpus (from English to Spanish) were used for translation quality assessment. From each corpus, a number of occurrences representing the population were sampled. In the first case study, after the occurrences obtained from the monolingual corpora had been classified and analyzed, quantifiers expressing 56 different functions were determined. In the second case study, the ten most frequently used nouns in each language were extracted from the bilingual corpus. A total of 7882 concordances were determined. In the final case study, identification of translation problems was realized through comparison of translation occurrences of English simple past into the Spanish preterite. After the input items had been searched for, 397 cases of the English simple pasts and 392 cases of Spanish preterite were detected. In the first phase of the study, data obtained from the two monolingual corpora were compared for similarities and differences in formal actualizations of specific semantic functions. In the second phase of the study, data obtained from the parallel corpus were compared in terms of the actualization of particular formal structures. Afterwards, the results were analyzed for meaning in order to see how the translated usage was distributed. In the final phase, data from the original Spanish corpus were compared with data of the bilingual corpus (translated Spanish only). Through this intralinguistic comparison, differences between the original and translated Spanish regarding the use of grammatical structures were determined. Analysis of quantifiers indicated that translators mostly used high-frequency Spanish words in the expression of the same function in the target language. This trend shows the use of a homogeneous type of language rather than L1 interference. Nominal characterization analysis revealed the influence of the source language on translations. Corpus based past tense analysis revealed that translated texts had a lower degree of accuracy and communicative economy than the original ones.

As the studies reviewed above have indicated, even though translation tasks have for years been included in language assessment, little effort has been made to compare and contrast the directly written and translated papers of students in order to provide more consistent writing assessment. However, although some of the studies discussed above did not specifically aim to measure the outcomes of translation used as a language task in FL classrooms, the results

obtained from them have provided valuable empirical evidence regarding the impacts of translation use in pedagogical contexts.

Studies on reliability and validity of translation tests.

The studies in the field of Translation Studies that examine the reliability and validity of translation tests are also limited in number (e.g., Colina, 2008; El-Banna, 1993; Eyckman, Anckaert & Segers, 2009; Ghonsooly, 1993; Ito, 2004; Neves, 2002; Orozco, 2000; Tavakoli, Shafiei & Hatam, 2012).

For example, focusing on examining the validity and reliability of translation tests as a reading comprehension measure, Ito (2004) compared open-ended translation tests with multiple-choice translation tests in terms of their effectiveness in improving learners' reading comprehension skills. Also, the researcher attempted to determine whether the language proficiency levels of the students had an impact on the validity and reliability of the tests. In total, 35 male and 35 female Japanese high school students with a similar language background and educational level participated in the experimental study. An open-ended translation test (from English to Japanese), a multiple-choice translation test, an open-ended cloze test and a short-answer question test were used as the data collection instruments. These tests were formed through the adaptation of a selected low-intermediate level reading passage (containing 317 words) by the researcher. The readability of the passage was measured by using the Flesch-Kincaid readability formula. The participants in the study were divided into four groups and they were given the tests in different orders. Administration of the tests was performed at the beginning of English classes by an instructor and an assistant. Completed test papers were rated by two Japanese English language teachers from a state university. The inter-rater reliability coefficients were measured as between .83 and .90. The findings of the study indicated the following: 1) in the measurement of reading comprehension, the open-ended translation test was found to be a reliable and valid method, 2) the multiple-choice translation test was found to be a reliable but moderately valid method, 3) a moderate correlation was detected between these two types of translation tests, 4) the performances of the groups were determined as similar across both of the translation tests, 5) proficiency level of the students was found to be an effective factor in the reliability and validity of the tests.

As translation testing methodology has always been considered to have a subjective nature, in his study Ghonsooly (1993) compared a translation test with a standardized test in terms of its validity and reliability as an assessment tool. In the adaptation process of the translation test, the researcher followed some procedures. First, the researcher selected some sentences from among scientific and technical English texts, since paragraphs were thought to

be too long for use in translation tests. Secondly, in the process of sentence type selection, the researcher set some criteria to be met such as discourse, frequency of occurrence of the sentences in academic contexts and their functions or sub-functions. Thirdly, two phases of pretesting were performed: in the first phase, a preliminary version of the test containing 40 items in an open-ended form was administered to 825 volunteer Engineering students. In the second phase, the same test was given to two translation experts. Then students' answers were compared with those of the experts in order to constitute a multiple-choice translation test and to ensure its objectivity. Distracters for each item were selected from the responses of the students according to their high frequency of occurrence. After the completion of the test, the multiple-choice translation test including 50 items was administered to another randomly chosen 55 students majoring in the Electronics Department of a state university. Then, applying standard item analysis technique, the researcher excluded 30 items from the original 50-item test. Lastly, the final version of the translation test was administered to 60 volunteer Engineering students. The reliability of the test was determined as .74. The concurrent validity of the test was found to be low and not significant. Factor analysis of the test showed that the translation test could be used in measuring students' comprehension skills and some other basic language skills such as grammar and vocabulary.

Eyckman *et al.* (2009) compared a norm-referenced method (The Calibration of Dichotomous Items) that they developed for translation assessment with two other methods (analytical and holistic) in terms of their reliability. The participants in the study were 113 university students attending Dutch-French translation courses at four different francophone translation colleges in Belgium. The ages of the participants ranged between 18 and 25. As a translation assignment, a short brief (consisting of 346 words) was given to the participants in their class hours and they were required to translate the passage into French. The brief was a non-specialized text of average linguistic complexity, written in a journalistic style. The translation task was carried out with paper and pen and learners were given 90 minutes to complete the assignment. Also, they were allowed to use manual and electronic dictionaries. The whole procedure was supervised by one of the researchers. Four professional raters (two male and two female) were recruited to score the translations of the students according to the method they preferred to use. While two of the raters evaluated the papers holistically, the others used analytic grids for assessment. In the process of test calibration, two of the researchers worked together on identification of all the translation mistakes. A total of 170 translation errors were determined, and they were listed according to a matrix developed for these items. Then, students were assigned a 0 or 1 score in terms of their performance on each item. The reliability of the test was calculated and it was found to have .958 Cronbach Alpha

Coefficiency. However, correlation coefficients between the graders ranged between .635 and .780. In conclusion, the results of the holistic assessment revealed less inter-rater reliability than the results of the analytic grids.

Focusing on the disadvantages of traditional translation tests used in colleges, El-Banna (1993) carried out a quasi-experimental research project in order to develop a multiple-choice translation test which could be used as an assessment tool in determining the level of proficiency at translating texts of English as a Second Language (ESL) learners'. In the development process of the test the researcher followed some steps. First, the sentences and passages that were to be used in the test were selected from a book 'Everybody's Conversational English'. All of the selected passages and sentences were related to daily life situations and activities like shopping, health, amusement, news, politics, etc.... Then, a committee of language experts evaluated the difficulty level of these sentences and passages, and they recommended some revisions and adaptations. From 55 passages, 15 sentences/passages were selected to be used in the first part of the translation test (from English to Arabic) and ten sentences/passages were chosen for use in the second part of the translation test (from Arabic to English). Subsequently, this version of the test was administered to 240 ESL freshmen students (120 of them answered the first part of the test and the other half responded to the second part of the test) in order to detect possible problems related to the selected passages/sentences and to generate distracters. The pilot test was scored by the researcher and the final version of the test (composed of two parts, each including 30 items) was constructed. Then, the test was given to another 92 freshmen college students. The intrinsic validity of the test was found to be .82. The reliability coefficient of the first part of the test was determined as .66 and second part's was found as .67, indicating that the multiple-choice translation test was good at measuring learners' translation skill and proficiency level.

Orozco (2000) conducted a research project, concerning with translation from L2 into L1, in order to assess first year translation students' acquisition of translation competence. First, dependent and independent variables (translation errors, translation problems and translation notions) were tested using three distinct experimental tools. Afterwards, validation and standardization processes for each test were carried out and all of them were found to be reliable. Each of the tests was administered to translation students from three state universities at two different times (one at the beginning of the first semester and the other after an eight-month interval). After a series of pilot studies had been performed, the researcher composed a single instrument measuring all three of these elements. The testing tool included 12 items regarding notions of translation and translation competence, and a section for problems of

translation, in which students were required to translate an originally English text into Spanish. In the evaluation of the text that students translated, normal error assessment procedure was followed: two points deducted from a possible 20 points for each major error. Analysis of the translations revealed four major problems: (1) grammatical problem, (2) extra-linguistic problem, (3) transfer problem, and (4) pragmatic problem. Orozco suggested that the instrument used in the study might be helpful for translation teachers in investigating and gaining insights into translation competence acquisition process of their students.

Neves (2002) carried out an experiment in order to investigate the reliability of a quality scale developed specially for translation assessment. For quality assessment, six undergraduate students attending translation courses in Brazil and six professional translators were asked to translate an English text into Portuguese. The whole translation process was recorded by Translog program and participants were not allowed to use any materials as reference except for the program's built-in dictionary. No time limits were set for the completion of the translation task. After the translations had been produced, they were given to four external evaluators with the assessment scale. The scale incorporated ten questions related to translation quality which were obtained from a scale developed for measuring writing quality assessment (Ransdell & Levy, 1996). For each aspect of translation quality represented by a question in the scale, the evaluators were required to assign points from 1 (not at all) to 5 (completely). The raters were not provided with information about the identities of the translators in order to eliminate any systematic influence. As a result of the study, the scale was determined to be reliable with .95 Cronbach Alpha coefficient. However, according to the evaluators, the scale was not sufficient for measuring the construct. Regarding the consensus between the scores appointed by each rater to each question in the scale, highly significant correlations were determined.

In another empirical study, Tavakoli *et al.* (2012) attempted to examine the reliability and validity of translation tests seeking to prove that translation can be applied as a reliable assessment method for measuring reading comprehension level of learners. To achieve this objective, initially, the researchers compared an open-ended translation test with a multiple-choice translation test. Then, these two tests were contrasted with a multiple-choice reading comprehension test and an open-ended cloze test. In total, 40 Persian students studying English Literature and Translation at a state university in Iran participated in the study. All of the aforementioned tests were constructed from a reading book titled 'Book 10: A Course in English'. The readability level of the reading passages used in the tests was measured as 8.2 by application of the Flesch-Kincaid Grade level test, and it was determined to be suitable for the

proficiency level of the students. Each of the tests was administered to the participants in a different session. Two of the researchers from the study scored the results of the four test papers. Also, the researchers set a benchmark in order to assess the translation tests objectively. A number of statistical tests were performed; the open-ended translation test's reliability coefficient was found to be .87 and its validity was determined as .89. The multiple-choice translation test was found to be reliable (.79) and moderately valid (.69). Therefore, it was determined that the open-ended translation test was more reliable and valid than the multiple-choice translation test. Further, a high positive correlation was found between the multiple-choice reading comprehension test and the open-ended translation test (.89). Similarly, a high correlation was found between the cloze-test and the open-ended translation test (.79). Consequently, the findings of the study showed that the open-ended translation test could be applied as a reliable and valid testing method to assess the reading comprehension skill of students.

Taking a functionalist/componential approach, Colina (2008) developed a translation quality assessment tool. In order to provide some empirical evidence for the proposed approach and to measure the reliability of the tool, the researcher conducted a testing study. The design of the tool, evaluation criteria and descriptions of the components were all based on the theoretical principles of the functionalist approach. The tool incorporated four evaluation categories: specialized content, non-specialized content, functional/textual adequacy and terminology, each of which included a number of descriptors as sub-components. A numerical value was assigned to each of the descriptors. These varying values were used in the calculation of overall translation quality. Before the actual testing of the tool, the researcher conducted a pilot session with a number of raters and it was proved to be reliable. The actual testing session was held two weeks after the pilot testing. A total of 40 raters with three different L1 backgrounds (Chinese, Spanish or Russian) participated in the study. The raters were language teachers, professional translators and bilinguals. The texts intended to be scored by the raters were randomly selected translated Public Health materials retrieved from the web sites of federal agencies (e.g. Centers for Disease Control). Before the beginning of the rating procedure, the raters were given information about the tool and they were trained in how to use it. Then they were given the translated texts along with the briefs and asked to assign points to each of the texts. In the calculation of inter-rater reliability, the Excel 20. Program was used. The results of the study indicated good inter-rater reliability: by text .8 and by language group .7.

Studies conducted on the use of translation as an assessment tool.

Since translation has a complex nature, it has not been given enough attention by researchers in respect of its use as an assessment tool in an academic setting. Thus, the number of the studies conducted on this issue is limited in the literature (e.g., Calis & Dikilitas, 2012; Kallkwist, 1988; Laufer & Girsai, 2008; Lee, 2013; Prince, 1996)

In a Turkish context, Calis and Dikilitas (2012) conducted a classroom-based research to understand elementary level English learners' reactions to the integration of translation tasks into L2 learning practice. The participants in the study were 28 elementary level preparatory school students learning English for their academic education. These students were taught some predetermined grammatical structures through the use of translation practices for about 7 weeks. The data of the study were collected through the administration of two sets of questionnaires (The Inventory for Beliefs about Translation and The Inventory for Translation as a Learning Strategy). The results of the study showed that the use of translation tasks and exercises increased learners' reading comprehension and helped them memorize target vocabulary. In conclusion, translation tasks were found to be efficient in promoting learners' receptive and productive language skills.

In a study carried out by Lee (2013), the impact of the use of translation tasks on L2 learners' reading comprehension skills was examined. In total, 35 students attending a Chinese/English translation course at a university in Taiwan participated in the study. All of the participants had more than seven years of English learning experience. The researcher assigned the students to two different groups: an English majors group and a non-English majors group. The steps in the procedure were as follows. First, both of the groups were given an originally English passage composed of seven paragraphs. Secondly, they were asked to read the whole passage for about 15 to 20 minutes. Then, they were asked to translate the first two paragraphs of it into their mother tongue (Chinese). After that, they were required to respond to five multiple-choice reading comprehension questions. Three of the questions were related to the paragraphs that students translated and two of them were inference questions. For the assessment of the translations, two criteria were applied: accuracy and expression. In order to grade the translations, the researcher applied professional translation assessment. For accuracy, different points were appointed to the translated texts regarding their semantical equivalence or resemblance to the original text. For expression, readability and understandability features were considered. After the assessment, the scores that students got from the translation task were compared to the scores they got from the reading comprehension test. After a while, another passage was given to the students and the same procedure was followed. The passages used in

the study were extracted from a textbook by the researcher and they were selected due to their relevance to the daily lives of university students (becoming adults and love relationships). The findings of the study indicated that both English majors and non-English majors gave more correct responses to the questions related to the parts of the passages they translated. Thus, the findings of the study showed that translation had a positive impact on students' reading comprehension skills. It was implied that the incorporation of translation into language classrooms can be helpful in improving students' language proficiency.

In order to determine whether translation can be applied as an assessment tool in defining general proficiency levels of FL learners, Kallkwist (1998) compared the performance of English as a Foreign Language (EFL) students in translation and free essay writing. A total of 197 adult Swedish learners majoring in the English Language Department at Lund University took part in the study. These participants whose ages varied from 19 to 48 years old had similar prior English learning experience (9-10 years). The data of the study were collected from three sources: translation tests from L1 into L2, recall of a narrative and free essay. Free essays were completed by 17 students and 30 students volunteered to take the written retelling task. The rest of the participants were administered three translation tests (each of which was completed by 50 students). Thirteen native speaker judges and a committee of lecturers carried out a detailed error identification procedure by providing consistent and careful judgments regarding lexical errors found in the writings. In the error identification process, the judges underlined each infelicitously used lexical item and provided suitable alternatives for them. The findings of the study showed that the infelicitous use of lexical verbs was found more often in the retelling and composition data than it was in the translation tests. The researcher suggested that the reason why the results for translations differed from those of the two other data sets might be based on translation learners' being unable to resort to avoidance strategy.

Focusing on the superiority of translation activities in enabling learners to remember target words, Prince (1996) compared contexts of non-translation and translation in terms of their outcomes. Forty-eight students from the Pharmacy Faculty of the University of Montpellier participated in the study. After the students had been randomly assigned into two groups (translation group and non-translation group), the experiment was carried out in two phases. In the first phase of the study, the translation group was taught 44 target words along with their translated equivalences, the non-translation group was exposed to target words given in L2 contexts. The second phase of the study was a recall session held after a 40-minute interval. All the participants were given a paper with 22 target words on it to translate and another paper including 22 English sentences in which previously taught target words were

embedded was administered to them as a gap-filling task. The performance that the students displayed in the recall session was assessed in terms of the percentage of the accurate responses by condition of learning. The results presented by group indicated that translation instruction was significantly superior to context instruction in respect of the quantity of new target words retained. In this study, contrary to the assumption that the use of L1 interferes and thus hinders L2 learning, the use of translation seemed to bring benefits to the acquisition of target language vocabulary.

Another investigation focusing on the impact of translation on the acquisition of L2 vocabulary was undertaken by Laufer and Girsai (2008). Seventy-five high school students (aged 15-16) participated in the study. All of the students were native Hebrew speakers who had intermediate level English language proficiency. The researchers randomly assigned them into three groups and they were provided with three different instructional methods on L2 vocabulary acquisition: (1) form-focused instruction with focus on the target words, (2) meaning-focused instruction with no focus on the target words, and (3) translation and contrastive analysis with focus on translation from L2 into L1, and vice versa. Ten collocations of verb-noun and ten unfamiliar single words embedded in a reading text were used as the target items. In the first stage of the experiment, all three of the groups were exposed to identical procedures. All the participants were asked to respond to 13 true-false statements based on a reading comprehension test administered to them. For this task, no dictionaries were allowed and the meaning of the unfamiliar words was provided by the teacher. In the next stage, two communicative tasks related to the previously given reading text (open-ended reading comprehension questions and fill-in-the-blank exercises) were administered to the meaning-focused group. The form-focused group was given two form-focused tasks (a multiple-choice vocabulary test and a gap-filling exercise based on single words embedded in the same text). The translation group was required to complete two sentence-based translation tasks (from L1 into L2 and vice versa). Following that the target items were contrasted explicitly with their L1 equivalences. After a week, all the participants were tested to determine whether they were able recall the target words both actively and passively. While passive retention was assessed by the ability of the participants to translate words from L2 into L1 and provide definitions for them in English, active retention was determined by their ability to provide target words in reply to their L1 translations. The findings of the study revealed that the translation group significantly outperformed the other groups on all of the tests. Laufer and Girsai state that the cross-linguistic instruction improved acquisition by enhancing the salience of the target forms.

Studies on the impact of translation on L2 learners' accuracy.

Most of the studies in the field of Translation Studies reveal the positive impact of translation use on L2 learners' grammatical accuracy (e.g. Berggren, 1972; Ghaiyoomian & Zarei, 2015; Kallkvist, 2008; Soleimani & Heidarikia, 2017; Uzava, 1996; Vaezi & Mirzaei, 2007). However, a few studies have asserted that L1 interference in translation lead students to make more mistakes (e.g. Schjoldager, 2003).

Marie Kallkvist (2008) carried out a longitudinal, experimental study at Halmstad University, Sweden in order to determine the impacts of translation as a focus-on-formS activity on learners' L2 morphosyntactic accuracy. In total, 55 native Swedish-speaking students whose ages ranged from 18 to 37 volunteered to take part in the study. All of the students had at least nine or ten years of English language learning experience. Before implementation of the study, the students were given a form-focused multiple-choice test of English grammar and according to the results of the test they were assigned (by following a matched-pair random assignment) into two treatment groups: Translation group (T) and No translation group (NoT). Later, the researcher included a third group in the study in order to compare the treatments of the focus-on-formS exercise with a treatment of focus-on-meaning. The students who comprised this group had nearly the same level of English language proficiency as the students in the two experimental groups and they were attending an elective English course entailing reading fiction in English. This third group was referred to as the No grammar (NoG) group, as they were not given explicit English grammar instruction. Both of the experimental groups were given 13 weeks of explicit grammar instruction by the researcher. While the first group was engaged in translation exercises related to the target structures, the second group was given transformation and fill-in-the-blank exercises regarding the same target structures. In order to investigate the classroom discourse that resulted from the two different types of exercise, the researcher audio-recorded most of the lessons. All of the students who participated in the study were administered a multiple-choice test, a translation test and a written retelling task as pre-tests and post-tests. After the implementation of the first two tests, all of the students were given a Swedish text containing 21 target structure tokens and they were required to write down what they understood from the given text in detail and in good English (the same text was used both in pre-test and post-test). The results obtained from the pre-tests revealed no significant differences among the three groups. In the post-tests, in the multiple-choice and translation tests, T group outperformed the NoT and NoG groups and both the T and NoT groups had higher gain scores than the NoG group. However, in written retelling task, the NoT group performed superior to the T group and both the NoT and T groups performed significantly better

than the NoG group. Further, in a comparison of the instruction types, students who received explicit instruction on target structures were found to perform better than students treated with implicit grammar instruction.

In an exploratory study focusing on lexical matters, Berggren (1972) examined the use of 'eight words and expressions' in English by 120 Swedish undergraduate students of English, before and after 45 minutes of instruction. All the subjects involved in the study were advanced learners with approximately eight years of previous English learning experience. The participants were randomly assigned into 16 groups. Then, these groups were divided into two and each half was taught using a different type of instruction based on a brief text. The first eight groups (composed of 57 students) were given a text to translate from Swedish into English and the other eight groups (including 63 students) were asked to carry out close reading of the same text in English. The text used in the study was originally constructed based on two native English speakers' descriptions of a picture series. The total time allocated for each instruction session was 45 minutes. The teachers who were involved in the instruction process were provided with detailed information about the phrases, words and structures to be discussed with the students during the lessons. In this way, all of the students in the groups were exposed to the same words and expressions, but only the students who translated the text contrasted them with their equivalents in Swedish. Before the beginning of the 45-minute period of instruction, all of the participants were required to write down a detailed description of the given picture series. A week later (after the 45-minute instruction session), the students were asked to describe the same picture series again in order to measure possible progress changes between the first and second attempts at the same task. The results of the study revealed that the number of interference errors was higher in the close reading groups than in the translation groups. Thus, it was concluded that the use of translation had a positive impact on L2 learners' accuracy.

In order to understand whether the use of translation has an impact on the number of the errors that L2 learners make, Schjoldager (2003) conducted a comparative analysis of errors that occurred in L2 picture verbalizations and L2 translations. The subjects of the research were 21 third-year secondary school students who were specializing in language at RisikovAntsgymnasium and 32 third-year university students majoring in English Language Education at Aarhus University. The mother tongue of all the students who participated in the study was Danish and all of them had previously taken English as a FL lessons. These students were randomly assigned into two groups: a translation group (composed of 10 secondary and 16 university students) and a picture verbalization group (comprising 11 secondary and 16 university students). For comparison, the researcher selected translations and picture

verbalizations, in particular, as they were both controlled tasks that required students to put their ideas into words. The pictures used for the verbalization task were obtained from a children's book by Pia Thavlov (1996), including many descriptive details about the life of a family. The source text that was to be given to the students for the translation task was constructed based on four Danish native speakers' verbalizations (426 words) of the same story pictures. Both of the groups were provided with an adequate time limit (65-75 minutes) for completion of the tasks. As a result of the study, it was determined that the students in the translation group made more errors than those in the picture verbalization group. While the translation group made 709 errors the picture verbalization group made 425 errors. The researcher claimed that enhanced L1 interference in translation caused students to make more mistakes.

In another study conducted by Soleimani and Heidarikia (2017), the impact of literal translation (used as a cognitive strategy practice) on learners' acquisition of particular English grammatical structures (participial phrases and absolute constructions) was examined. In total, 40 Iranian EFL university level students (11 males and 29 females) participated in the study. After a pre-test had been administered to the students for the purpose of determining their prior knowledge regarding the targeted language structures, they were randomly assigned to one experimental group and one control group. For treatment, the experimental group was provided with 10 English sentences including targeted structures and they were asked to think aloud while verbally translating them into Persian. The session was held during the regular class time of the students and the whole procedure was recorded by one of the researchers. On the other hand, the control group was given deductive instruction in the targeted structures. Both participial phrases and absolute constructions were taught to the participants of the control group by an instructor and one of the researchers. The students were required to work on some of the exercises in the course book. The data of the study were collected over pre- and post-tests, each including 30 multiple-choice questions related to the targeted structures of the study. The students were given 30 minutes to complete the tests and each correct answer was worth 1 point. In the analysis of the data, parametric statistics were applied. The results of the study indicated a significant difference between the experimental and control groups. In the post test, the participants in the experimental group got significantly higher scores than the participants in the control group. The researchers argued that this could be a result of translation's dramatic effect on students' awareness of the focus on grammatical structures.

In 1996, focusing on processes of writing and translation, Uzawa investigated ESL learners' attention patterns and the quality of the language that they used. Twenty two Japanese

ESL learners from a Canadian college participated in the study. After the participants had been trained in thinking aloud, three tasks (including writing a draft of an essay in L1, writing an essay on a given topic in L2, and translating a brief article from L1 into L2) were administered to them in a randomized order. The researcher recorded each task administration process and took observational notes. Analysis of the data was performed by two trained experts. In the analysis of the data obtained from think aloud protocols (TAPs), the researcher addressed the meta-cognitive, linguistic and discourse level attention of the learners and their personal comments. Observational notes were used to support protocol analysis. After the evaluation of the writing samples had been completed, the researcher concluded that students mostly applied a 'what next' approach in the tasks in L1 and L2. In the translation task most of them followed a 'sentence by sentence' approach. Further, it was determined that in the L1 and L2 tasks students adopted similar attention patterns, whereas in the translation task they paid significantly higher attention to language use. In the L1 and L2 writing tasks, students cared about organization and what to write next. On the other hand, in the translation task, they paid attention to using language more accurately and facilitated from a wide range of vocabulary. Thus, Uzawa asserted that translation provides conscious L2 learning which coincides with what some other researchers have claimed (e.g., Ellis, 1986; Mclaughlin, 1987).

In another study, Vaezi and Mirzaei (2007) investigated whether the use of translation from L1 to L2 has a positive impact on EFL learners' linguistic accuracy. The participants in the study were 72 pre-intermediate level Persian EFL learners (who were divided into two groups: a control group and an experimental group). Four structures were taught to both of the groups over 16 sessions: indirect reported speech, the second conditional and wish, followed by the simple past tense and the passive voice. After the introduction of each structure, the experimental group was given 24 sentences to translate individually in class (six per session). On the other hand, the control group was asked to complete six grammatical exercises (per session) obtained from a course book. At the end of each session, the correct answers were provided by the instructor and further explanations in terms of grammatical structures were given. The results of the post-treatment test revealed that the translation group had a significantly higher mean score than the exercise group. Translating from L1 to L2 was found to be useful in producing new structures and forms. Thus, in this study, translation was found to be an effective task regarding linguistic accuracy. Vaezi and Mirzaei claim that if learners are made aware of the differences between L1 and L2 structures, they can easily avoid negative interference from mother tongue to the language being learned.

Ghaiyoomian and Zarei (2015) carried out an experimental study in order to examine the impact of using translation on the grammatical accuracy of EFL students. To fulfill this purpose, 62 male junior high school students were included in the study. In the selection of the participants, a researcher-made pre-test including a number of items for testing three grammatical structures (adjectives, possessives and word order) was used. In this way, students who were not familiar with the intended structures were determined. After the pre-test, the students were randomly assigned into an experimental group and a control group. Whereas the students in the experimental group were provided with some sentences (related to the intended structures) to translate from Persian to English, the control group just did the exercises in their textbook. At the end, both of the groups were administered a post-test which was similar to the pre-test in terms of format and content. In the analysis of the data descriptive statistics were used. The post-test results revealed that students in the experimental group who were treated with a form of translation got significantly higher scores than the students in the control group. Thus, in this study, the use of translation was found to be an effective practice in improving EFL learners' linguistic accuracy.

In line with most of the theoretical arguments suggested in favor of pedagogical translation, the empiric research reviewed thus far has yielded mostly convergent results that strongly support the re-examination of translation as a cognitive tool facilitating the acquisition of language by testing learners' linguistic competence. Therefore, it is reasonable to claim that translation can be used as a toolkit by language teachers since it improves L2 learners' grammatical accuracy and helps them develop a wide range of skills throughout the L2 acquisition process (Laviosa, 2014, p. 34).

Summary and Research Gaps in Translation Assessment

In brief, this chapter touched upon the issues of translation quality, subjectivity, reliability, validity and objectiveness in translation quality assessment and discussed some considerations including major approaches to and models for assessing translations. The chapter continued with a review of a number of empirical studies conducted on the impact of translation on L2 acquisition and accuracy; on the comparison of translation tasks with other monolingual and bilingual tasks; and on the examination of the validity and reliability of translation tests.

In spite of the fact that a significant number of studies have been carried out on translation evaluation over decades, there is still little agreement in the translation assessment field, and there is a lack of reliability in assessment tools and a lack of adequate rigor in judging quality (Marais, 2013, p. 13). In the literature, limited research has been conducted to examine

the impact of rating experience on translation score variability. In addition, to the best knowledge of the author, there have been no studies conducted on the issue of the reliability of scores in translation assessments in the Turkish context. Thus, this thesis study aims to fill this gap in the literature by examining issues of rater reliability in EFL translation assessment in the context of higher education. The findings obtained from this study will provide insights for practices of assessment at the institutional level. Also, in this study, the impact of translation quality on translation scores variability has been investigated. Thus, the results are assumed to shed light on the issues of translation quality in relation to the assigned scores, interaction among evaluators and their decision-making processes.

Studies conducted in the field of translation assessment have applied various quantitative and qualitative research methodologies: mathematical/statistical measures to identify the value of a translation or of the responses of the readers; and questionnaires and interviews for the qualitative evaluation of the quality of a translated text (Yamini & Abdi, 2010). However, the use of TAPs has not been widely preferred by the researchers since it is a challenging task for them to collect, prepare (transcription) and analyze verbal data. Therefore, this thesis study is assumed to contribute to the translation assessment literature by investigating rater cognition through verbal protocols. Additionally, to the best knowledge of the author, no studies in the Turkish context have dealt with the issue of the decision-making behaviors of raters during the process of scoring translation papers, which constitutes a research gap that this study aims to fill.

CHAPTER THREE

Methodology

The aim of this study is to measure the reliability and objectivity of the raters' scores given to the scripts that EFL students translated in pedagogical contexts and to make suggestions about the elimination of the main affective factors that have an impact on the translation evaluation process. Applying a research design of mixed-methods, the data of the study were gathered both quantitatively and qualitatively. In the social sciences, applying a mixture of qualitative and quantitative data collection procedures has its own advantages. While quantitative experiments and surveys include procedures such as collecting data, analyzing them and reporting the results, qualitative research comprises case studies, G-theory studies and ethnographies (Creswell & Plano Clark, 2011). In a mixed methods research design, both qualitative and quantitative procedures are followed in a single study and the application of multiple methods strengthens the results obtained from a study by providing in-depth insights into the problem and by neutralizing the disadvantages of a particular method (Creswell, Plano Clark, Gutmann, & Hanson, 2003). According to Creswell *et al.*;

“A mixed methods study involves the collection or analysis of both quantitative and/or qualitative data in a single study in which the data are collected concurrently or sequentially, are given a priority, and involve the integration of the data at one or more stages in the process of research.” (2003, p. 165).

In the literature, various names have been given to the procedure of mixing research methods such as triangulation studies, mixed research, multimethod, multiple methods and blended research (Harrison, 2013). As a mixed methods research design, this study specifically applied convergent parallel design (Figure 3). In this research design, both quantitative and qualitative data are collected simultaneously; however, analyses of both data strands are kept separate, then the databases are mixed for overall interpretation of the results (Creswell & Plano Clark, 2011). Convergent parallel designs are mostly conducted in order to compare and contrast the results obtained from both qualitative and quantitative data strands and to confirm, validate or strengthen the quantitative results with the qualitative results (Harrison, 2013). In this design, quantitative and qualitative research methods are given equal priority in addressing the main research problem (Cresswell, 2011, p. 541).

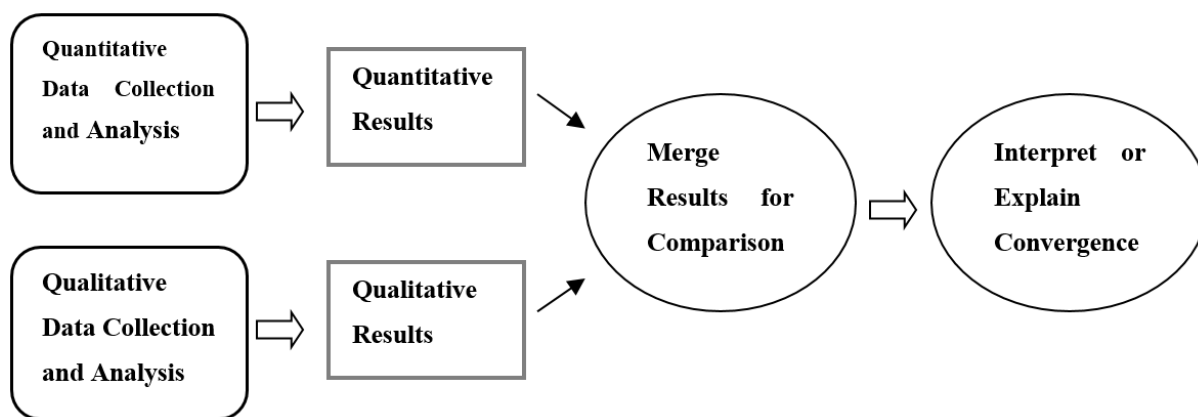


Figure 3. Diagram of the applied convergent parallel design (Creswell & Plano Clark, 2015, p. 56).

In this research, while comparison of the variability of the scores awarded by the raters to the same translation products constitutes the quantitative dimension of the study, the data obtained from the voice recordings (TAPs), through which the raters expressed their opinions, comments and perspectives in the process of scoring, constitute the qualitative dimension of the study. In the study, quantitative and qualitative data were collected simultaneously; however, the analyses of the data sets were carried out separately. The findings were then interpreted.

In the beginning of this chapter, the theoretical framework and the participating students and raters are described. Then, it provides detailed descriptions of the data collection instruments. The following sections describe the data collection procedures. Then, the steps for data analysis are described in detail. Lastly, an overview of research ethics and a summary of the section are provided.

Theoretical Framework

The theoretical frameworks that guide this translation assessment research are: Dynamic Equivalence Theory, Constructive Alignment and the G- theory approach (Elorbany & Huang, 2012). These theories will be described in detail in the following section.

Dynamic equivalence theory.

Equivalence is one of the key concepts in the field of translation. In translation theory and practice, the whole written corpus concentrates on equivalence since it is regarded as the only reliable measure for translation adequacy (Barghout, 1990). According to Nida equivalence is the phenomenon of generating a language that sounds naturally equivalent to the message of the source language regarding both its meaning and style (1964, p.19). Nida defines two different types of equivalence: formal equivalence and dynamic equivalence. In formal

equivalence, the translator pays particular attention to the linguistic form of the source text. This can be labeled verbatim translation, in which the translator seeks sentence to sentence or concept to concept correspondences. On the other hand, in dynamic equivalence, the translator is essentially concerned with creating an 'equivalent effect' by reflecting the relationship between the message of the source text and its receivers in the translated text in the same manner (1964, p. 159).

According to Kelly, dynamic equivalence is the act of transporting a unit of the source text equivalently in communicative effect. Since languages have distinct lexes, there is no need to struggle for word by word translation. Thus, in order to produce an adequate translation of a source text, we should perceive it in its entirety (1979, p.132). At this point, the notion of 'natural' equivalence emerges. What is posited here is that a source text and its translation activate the same cultural function and have the same value. Thus, when translating from one language to another, it is important to obtain the necessary cultural information. For example, in English-language cultures 'Friday the 13th' is considered to be an unlucky day, but for other cultures this might not be the case. In Spanish for instance, 'Tuesday the 13th' is believed to bring bad luck. When translating from Spanish to English, we might either say 'Friday the 13th or Tuesday the 13th'. If we do not change the day, then it is called formal equivalence. However, if we change the day by paying regard to the cultural beliefs of that language community, then it is called dynamic equivalence (Pym, 2010, pp.139-140).

Nida argues that formal equivalence requires a translation to correspond with its original as much as possible regarding the aspects of both content and form. On the other hand, based upon the equivalent effect principle, dynamic translation necessitates the translator to apply some grammatical, lexical and cultural adaptations to create a relationship between the message and its receiver similar to the one between the original message and the receiver. Thus, by making some adjustments on culture related language forms, the translator avoids any interference stemming from the source language and minimizes the unfamiliarity with the source language setting (1964, p.159). Koller (1979), another theorist interested in the concept of equivalence, associates dynamic equivalence with communicative equivalence by emphasizing its relationship with the target text audience for whom the translation is intended (cited in Munday, 2001, p. 47). Influenced by Noam Chomsky's generative-transformational model, Nida (1964) suggested a system of a translation that consists of three main stages: analysis, transfer, and restructuring. In Nida's science of translation, a translator decodes the source text by analyzing its surface and deep structures and then by applying a technique of encoding, s/he transfers these elements into the target text in order to restructure it both

stylistically and semantically (Munday, 2016, p. 63). Figure 4. presents Nida's three-stage system of translation.

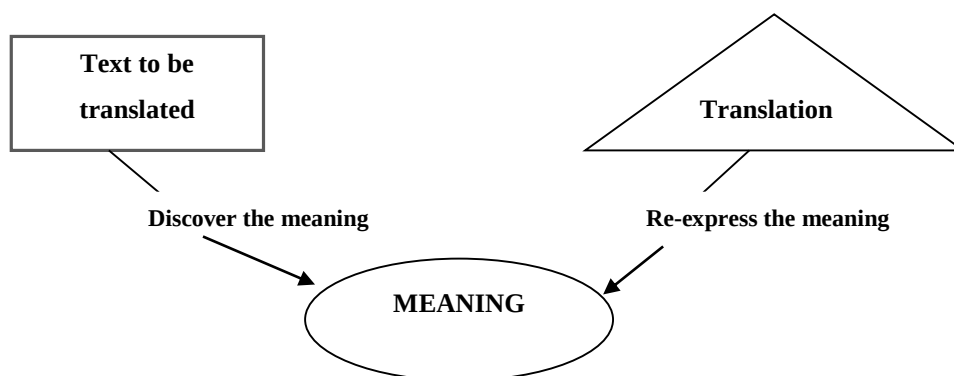


Figure 4. Nida's three-stage system of translation (Nida & Taber, 1969, p. 33).

In this thesis study, following a source-oriented translation quality assessment approach, the translated products of the students were evaluated according to Nida's dynamic equivalence theory, as the two texts (source text and target text) were compared with each other regarding aspects such as purpose, culture, target audience, text, content and language.

Constructive alignment: theoretical construct of the instrument used in the study.

As a well-known concept guiding the improvement of learning opportunities, especially at the tertiary level, constructive alignment provides a student-centered assessment by linking it to the quality of learning (Nightingale, Carew & Fung, 2007). In the constructive alignment framework, the learning outcomes that students are required to acquire are defined in advance of the teaching process (Biggs, 2014). Fundamental to this approach is the idea that every student constructs his/her learning in relation to the activities of learning (Biggs, 2003). According to Biggs (1999), assessment at tertiary level can be described as 'backwash'. Thus, it is essential to shape the way that students respond to instruction rather than to the assessment (Nightingale *et al.*, 2007). In this regard, constructive alignment suggests that if students are expected to achieve the learning outcomes efficiently, then the task assigned to them should incorporate learning activities that will facilitate their achievement of these outcomes (Brown, 2001). What is meant here is that teaching, strategies used for learning, curriculum, evaluation methods and instruments should all be aligned by reinforcing one another in order to attain the same intended outcomes (Marais, 2013).

Angelelli (2009) asserts that any assessment should be based upon an assessment construct. Depending on this construct, an assessor understands what is tested and which skills and competencies are required. Moreover, dividing these competencies into smaller units, defined previously in detail, helps the assessor mark a translation as either a good one or a poor

one. Following the principles of constructive alignment, Marais (2013) aligned the holistic rubric used in this thesis study with a particular theory (functionalist theory). Drawing on action theory, the functionalist approach deals with the active translation process, and it questions whether the translated product is successful in achieving ‘the intended communicative function’, which is the most essential criterion impacting on the translator’s decision-making process (Nord, 1997, p. 42). The functionalist theory suggests that the person who translates a source text into another language serves as a communication bridge between the members of societies from different cultures. Assuming a mediatory role, the translator adapts both verbal and non-verbal features of the source brief to the norms of the target culture. Thus, all the categories included in an assessment instrument are evaluated in accordance with the brief (Nord, 1997). Moreover, every brief has some requirements to be assessed such as purpose, culture, text type, language, style and accuracy. However, the adequacy of these features in a translation paper cannot be assessed separately as the quality of a translation is entirely based upon its wholeness (Marais, 2013).

According to Marais (2013), in translation assessment, the evaluation categories should be identified by the quality of understanding rather than counting the number of mistakes. Assigning scores to each of the mistakes in a translation paper cannot be regarded as a useful method of assessment since counting mistakes does not indicate the quality of understanding. The researcher advocates that a translated product should be handled as a whole. Namely, all the aspects of the translated work should meet the criteria determined beforehand. If one aspect in a translation fails to be fulfilled, then this translation cannot be marked as a good one. Also, the researcher created a checklist (see Appendix A) for the holistic rubric based upon Nord’s functionalist translation theory. This checklist was prepared to ensure that raters provided quick and consistent evaluation of the products. Furthermore, it was assumed to serve as a guide for the raters while they were commenting on the translated texts.

Thus, in this thesis study, the translated products of the students were assessed by following a criterion-referenced assessment with the use of holistic translation scoring rubric (developed by Marais, 2013). Also, the checklist was provided to the raters before the scoring process to facilitate their thinking aloud on the translated papers.

Generalizability theory.

In fields such as psychology, social sciences, and educational sciences, reliability is interpreted as the generalizability of the observed score of an individual to the average of the scores that the same individual gets in all possible circumstances (Güler, Uyanık & Teker, 2012). In other words, reliability is the degree of consistency or inconsistency of the

measurement results (Brennan, 2011). This definition of reliability is based on the assumption that an individual's knowledge, attitude, skill, or other measurable characteristics are in a steady state. In other words, it is assumed that any difference between the scores of an individual obtained at different times regarding the same measurement situation might have resulted from one or multiple sources of error, rather than from the individual's maturation or learning (Güler et. al., 2012).

When measuring a language competency, it should be noted that the construct that is to be assessed is totally abstract, and it is not possible to observe it directly. In other words, a true score that a person receives for a competency cannot be tested directly. Thus, the reliability of the scores obtained from a test should be based on the associations between the true score and the observed score (Bachman, 1990). A true score refers to the actual performance of an individual in a context of measurement and an observed score is obtained from the interactive relation between error score and the true score (Fulcher & Davidson, 2007). While a true score has two variances—error score and observed score variances—, an error score provides unsystematic and randomized variances that have no relations with true scores (Fulcher & Davidson, 2007). “Classical true score theory states that our observed score (X) is equal to the sum of our true score, or true underlying ability (T), plus the measurement error (E) associated with estimating our observed scores” (Schultz & Whitney, 2005, p. 69), or

$$X = T + E$$

where X , T , and E represent observed score, true score, and error score, respectively (Briesch, Swaminathan, Welsh, & Chafouleas, 2014). Although this equation is very simple and useful, there are some limitations that shadow these features. Classical theory is regarded as a weak testing theory accounting for only one variance source among multiple sources of error within a given analysis (Huang, 2012; Webb & Shavelson, 2005). Also, classical test theory fails in differentiating error, providing only a very general identification of probable and actual sources of measurement error (Webb & Shavelson, 2005).

Generalizability theory (developed by Cronbach, Gleser, Nanda and Rajaratnam) arose from the limitations of classical test theory (CTT) in 1972. In fields of social sciences such as psychology and education, generalizability theory, has presented a very comprehensive structure that calculates countless measurement situations by using variance analysis, which is a very powerful statistical procedure (Güler *et al.*, 2012). Regarding its conceptual framework, G-theory is different from CTT in several respects. First, while CCT accounts for only a single variance source, G-theory considers multiple sources of variability simultaneously. Second, G-theory predicts the magnitude of each of the variance sources. Third, G-theory calculates two

distinct coefficients of reliability relevant to decisions based upon the interpretation of both absolute score levels (Phi-coefficient) and relative levels of score (G-coefficient). While in absolute decisions, students' scores from an exam are used to classify them into certain groups, in relative decisions, exam scores are used to determine the ranking order of students. Finally, G-theory allows researchers to make decisions about reducing the impact of the variance error on the true score (Shavelson & Webb, 1991; Güler *et al.*, 2012). In comparison to CTT, which centers on estimating only a single error of measurement at a time (e.g. item, rater, form, etc.), G-theory enables researchers to analyze multiple sources of error variance simultaneously (e.g. raters, tasks, topics, etc.) (Brennan, 2001).

In examination of the reliability of behavioral measurements, G-theory ensures a very practical and flexible framework (Shavelson, Webb, & Rowley, 1989). The universe of generalization which is an essential concept in G-Theory is defined through multiple sources of error. In contrast to CTT, G-Theory defines true- and error-scores in multiple ways. While CTT is concerned with only two error sources 'a single ability and a single source of errors' (Bachman, 1990, p.188), superior to CTT, G-Theory defines every source of systematic and unsystematic error, distinguishes them and predicts each one of them (Webb & Shavelson, 2005). By extending CTT, G-Theory estimates the magnitude of multiple sources leading to measurement error (Shavelson *et al.*, 1989). For example, when two or more evaluators are asked to assign scores to a number of essays with different topics using either analytic or holistic scoring rubrics, some facets arise as sources that impact on the score variability. These facets can be identified as variability between essay topics, variability between rating methods (holistic and analytic), variability between raters and the interactive relationship among these facets. In G-theory, different variance sources such as rater, prompt, item and method of scoring are called facets and also this term 'facet' is used to make a distinction between the error sources and the factors in factor analysis (Briesch *et al.*, 2014). Also, source levels are addressed as conditions. For example, in situations where raters are treated as a facet, 1st rater, 2nd rater, 3rd rater, etc. are defined as conditions (Güler *et al.*, 2012).

Given the measurement scenario mentioned above, G-theory can predict the massiveness of the variance arising from each facet. Actually, when an individual's performance is measured, it provides a sample of behavior. Thus, the usefulness of a score is based to a considerable extent on accurate generalizability of a behavior in some wider contexts (specifically in a universe of generalization). In G-theory, prediction of the magnitude of facet variance is performed in two phases: a generalizability study (G-study) and a decision study (D-study). In a G-study, variance sources in test scores regarding each facet (occasion, rater,

instrument, student etc.) are first identified and then quantified. Thus, the necessary information can be obtained for the D-study to decide on individuals and individual groups (Huang, Han, Tavano, & Hairston, 2014). Also, the D-study enables the investigator to differentiate between the ultimate decision and the relative decision (Shavelson & Webb, 1991). In doing so, while the G-study examines the impacts of variance changes in the design of measurement, the D-study presents an integration of the design providing the interpretation of score reliability (Huang, 2008).

In the G-theory framework, apart from G-study and D-study analyses, there are some other fundamental concepts such as universe of generalization and universe of admissible observations (Brennan, 2011). While the universe of admissible observations is related to the conditions enabling a particular construct to be measured, the universe of generalization refers to a facet's conditions to which the results are generalized (Shavelson & Webb, 1991). In fact, while CTT engages in predicting a true score, G-theory centers upon the universe score which test-takers are anticipated to achieve (Briesch *et al.*, 2014). In generalization of an observed performance to universe score, there might occur some degree of error. For this reason, generalizability coefficients and dependability coefficients are used to calculate the extent of this error degree (Shavelson & Webb, 1991). In norm-referenced tests, the score that an examinee get is interpreted in relation to the performance of other test takers. Thus, in norm-referenced tests, generalizability coefficients are applied. On the other hand, in criterion referenced tests, in which the score that each examinee gets is interpreted regarding some previously defined test criteria, dependability coefficients are used (Brown, 2004).

In G-theory analysis, the design of the facets can be performed either *fully-crossed* or *nested* according to the way that the conditions of facets interact in the given design. Fully-crossed design is applied in situations where each of the conditions of a facet has interaction with other facets' conditions. But, nested design is used when only some conditions of a facet interact with other facets' conditions. To illustrate, in a case where a number of students responded to the same items and the answers given by the same students to the same items are rated by two different raters score items, a crossed design would be performed as follows: students by items by raters ($s \times i \times r$). However, in a situation where different students answered all the items in a test, and each student's answer is rated by a different rater, a nested design would be applied as follows: ($s : i : r$) (Güler, 2009). Also, in the G-theory framework, facets are distinguished either as random or fixed (Briesch *et al.*, 2014). When the conditions of a facet can be exchanged with the ones that exist in the universe, then the facet is regarded as

random. On the other hand, when the investigated instances do not require generalization, then the facet is considered as fixed (Briesch *et al.*, 2014).

In conclusion, by reinterpreting the classical theory of reliability, G-theory shows how the traditional discrimination between validity and reliability can be overcome in order to design dependable observations. Also, by determining multiple sources of error variance simultaneously and the impacts of their interaction, G-theory expands the use of CTT and provides a conceptual framework that compensates for the restrictiveness and weaknesses of CTT (Kieffer, 1998). With regard to the features provided by its theoretical framework and feasibility, G-theory is considered to be a more powerful and effective strategy than CTT in determining variability due to rater facet (Shavelson, Baxter, & Gao, 1993). Whether it is carried out in order to make an actual decision (criteria reference test) or to make a relative decision (norm reference test), G-theory provides a more integrated design for assessing reliability.

G-theory has been applied in the fields of psychology, education and social sciences for the development of performance assessment. And, it has been preferred by many researchers (e.g., Huang, 2008, 2011, 2012; Huang & Foote, 2010; Gebril, 2010; Schoonen, 2005) in the field of writing assessment to investigate validity and reliability of EFL/ESL writing scores. For example, Huang (2008) applied G-theory to investigate the reliability and variability of the scores given to native English essays and ESL essays. Similarly, Huang and Foote (2010) used the G-theory approach to investigate the variations of the scores assigned to ESL and native English writing papers. Using G-theory, Gebril (2010) examined the impacts of two different writing tasks and attempted to determine the effects of rater facet on the generalizability of scores. Huang (2012) conducted D-studies and G-studies to investigate the accuracy and construct validity differences of the scores given to ESL and native English writing papers. Also, Schoonen (2005) applied G-theory analysis to determine the impact of factors such as raters, writing aspects and rating methods contributing to variability in scorings.

As Webb and Shavelson claim “in order to evaluate the dependability of behavioral measurements, a generalizability (G) study is designed to isolate and estimate as many facets of measurement error as is reasonably and economically feasible” (2005, p. 717). In this study, it has been deemed appropriate to apply G-theory instead of CTT since G-theory can examine multiple measurement errors which might be due to raters, translation quality and translation papers (as determined facets in the present research design), and since it can account for measurement sources of error and the impacts of their interaction simultaneously. By applying this framework, the researcher could determine the variance sources and magnitude of them in

the score variability. Further, it enabled the researcher to optimise the conditions of measurement in the facets of the research.

Some application areas of generalizability theory.

G-theory has been applied by researchers particularly in three research fields: education, psychology and medicine. In psychology, G-theory is used to assess the reliability of individual job proficiency ratings (e.g., Kraiger & Teachout, 1991); to determine the reliability of test scores in measuring personal traits and individual differences (e.g., Vispoel, Morris & Kilinc, 2018); to determine the effects of measurement error on results (e.g., Lakes, 2013); to measure the reliability of ratings obtained from caregiver and child interactions (e.g., Lakes, Lucas, & Cooper, 2017); to measure the dependability of questionnaires (e.g., Li et al, 2018); to measure the generalizability degree of the data obtained from applied behavior analysis (e.g., Güler, Eroğlu, & Akbaba, 2014); to assess the extent to which measurements are contaminated (e.g., de Gruijter & Leo, 1991); and to estimate reliability for observational data (e.g., Lei, Smith, & 2007).

In the medical field, G-theory analyses are performed in clinical patient research settings to determine the reliability of measurements (e.g., Heitma, Kovalski & Pugh, 2009). Also, G-theory is used in clinical child and adolescent psychology research to determine the effects of measurement error on findings (e.g., Lakes & Hoyt, 2009) and to determine the construct validity of scales used to measure a personal trait (e.g., Lakes, 2012). Additionally, G-theory is used to develop performance assessment of the cognitive domain of children and to minimize measurement error in the field of medicine (e.g., Radzi, Nor, & Matore, 2018). Moreover, it is used to estimate the reliability of measurement instruments (e.g., Hall, 2015; Trejo-Mejía, Sánchez-Mendiola, Méndez-Ramírez, & Martínez-González, 2016). Furthermore, it is used to estimate sources of measurement error in order to improve the assessment of medical students (e.g., Tavakol & Brennan, 2013). In this field, the G-theory approach has also been applied to simulation performance evaluation (Prion, Gilbert, & Haerling, 2016). It is also applied by researchers to estimate the accuracy of a current measurement procedure (e.g., Oosterveld & Cate, 2004).

In the field of education, G-theory has been applied by researchers with the following aims: to demonstrate the relative effect of class size on the dependability of student ratings of instruction (e.g., İbrahim, 2011); to understand the decision-making behaviors of raters and to evaluate rater judgments (e.g., Şahan, 2018; Wang & Luo, 2009); to estimate the impact of rater experience on reliability of the writing scores (e.g., Han, 2013; Şahan, 2018); to measure the reliability, validity and dependability of scoring (e.g., Lee, 1999; Yin & Shavelson, 2008); to

estimate inter-rater reliability (Goodwin & Goodwin, 1991); to compare the reliability of scoring obtained from two different rating methods (e.g., Swartz *et al.*, 1999); to measure the individual differences of learners over personality tests (e.g., Tsutsui, Owada, Kondo, & Nakano, 2006); to investigate the reliability of self and peer assessments (e.g., Teker, Sahin, & Baydemir, 2016). It is also applied to reading assessment for students with cognitive disabilities (e.g., Tindal, Yovanoff, & Geller, 2010).

Selection of Raters

A total of 10 raters (4 females and 6 males) participated in the study. In the selection of the raters, a purposive convenience sampling method was applied according to the voluntariness of the participants to take part in the research and their proximity to the researcher. The raters who participated in the study had more than a year's experience in teaching and evaluating English Language and Education and had a bachelor's/master's/doctorate degree in English Language Teaching. All the raters were professionals in the field of interdisciplinary English language teaching, learning and assessment, and regular employees of the School of Foreign Languages (SFL), Foreign Languages (FL) Department, English Language and Literature (ELL) Department, or ELT Department at a state university in Turkey. They all had the same L1 background (Turkish). The raters had varied professional experience in teaching and assessing EFL writing tasks. Prior to the main data collection, an adapted rater profile form (Barkoui, 2008; Cumming *et al.*, 2001) was given to the raters to obtain information about their personal, educational and professional backgrounds (see Appendix B). Also, they were required to state their EFL writing assessment experience in years and which assessment method (holistic or analytic) they mostly applied while scoring writings. Two experience levels were determined (experienced and inexperienced groups) regarding the raters' years of experience in grading EFL translation products. Categorization of rater experience is shown in Table 2.

Table 2. *Rating Experience of the Participants*

		Experience rating EFL writing					Total
		2 years or less	3-4 years	5-6 years	7-10 years	11 and + years	
Experience Group	Inexperienced	2	2	1			5
	Experienced				1	4	5
Total		2	2	1	1	4	10

As is seen in Table 2, of the 10 raters 5 reported 6 years or less experience of rating the translation papers of EFL students. These participants were categorized as inexperienced raters. The remaining five raters declared that they had 7 years or more grading experience and they were categorized as experienced raters.

Table 3 illustrates the distribution of raters' gender and age based on the experience group.

Table 3. *Gender and Age Distribution of the Raters*

		Gender		Age		
		Male	Female	31-40	41-50	51 and +
Experience Group	Inexperienced	2	3	5		
	Experienced	4	1	3		2
Total		6	4	8		2

As Table 3 shows, 6 male and 4 female raters participated in the study. While the inexperienced group included 2 male and 3 female raters, there were 4 males and 1 female in the experienced rater group. Regarding the ages of the participants, 8 raters were between 31 and 40 years old, while two of the raters were over 51 years old. Regarding the participants' experience, Table 4 presents the distribution of raters' institutional affiliation and their academic titles.

Table 4. *Distribution of Participants' Institutional Affiliation and Academic Title*

		Academic title			Institution	
		Instructor	Assoc. Prof. Dr.	Prof.	Kafkas U.	Other
Experience Group	Inexperienced	5			3	2
	Experienced	4	1		4	1
Total		9	1		7	3

As Table 4 depicts, while 9 of the raters were employed as instructors at their universities, one of the raters was working as an associate professor doctor. While 5 of the instructors reported themselves to be inexperienced raters, the remaining 4 instructors and one associate professor doctor fell into the experienced group. Regarding their institutional affiliation, 7 of the raters were working at Kafkas University and the remaining 3 raters were employed at different universities. Table 5 shows the raters' highest educational qualification level and rater training background of them.

Table 5. *Raters' Highest Education Level and Previous Training on Assessment of Writing*

		Degree			Previous training	
		BA	MA	PHD	Yes	No
Experience Group	Inexperienced		5			5
	Experienced	2	2	1	3	2
Total		2	7	1	3	7

Regarding the highest education level that the participants had, Table 5 illustrates that 2 of the participants had a BA degree, 7 of them were graduates of MA and one of them was a PHD graduate. Additionally, 3 experienced raters declared that they had received training on assessing writing, while 2 experienced and 5 inexperienced raters reported that they had not received training in grading translation papers previously.

Selection of Translation Students

In this study, the researcher needed translated texts produced by students for the raters to assess. In the selection of the students, the researcher followed a purposive sampling method. The students who were selected to participate in the study were majoring in the English Language and Literature Department at Kafkas University where the researcher was working as a research assistant. The researcher focused on the 4th semester students (batch 2018/2019) as they had participated in three translation classes before the application: English-Turkish Translation I, Turkish-English Translation I and English-Turkish Translation II. At the time of the application, the students were attending another translation class (Turkish-English Translation II) in the department. Thus, the researcher considered them capable of translating well. From among 115 sophomore students, 40 volunteer students (30 females and 10 males) whose ages ranged from 19 to 46 were included in the study. These students were not administered a level test before the translation procedure as they had entered the department having achieved approximately same points in the university entrance exam. Also, all of the students had taken foundation courses in a preparatory class of the university for one year and, after taking a final examination, had earned the right to continue in the first class of the department. Therefore, they were all assumed to have nearly the same level of English language proficiency (upper-intermediate). Also, a written informed consent form was provided to the participating students affirming that they had the right to discontinue their participation at any time should they so wish (see Appendix C).

Data Collection Instruments

Following a mixed-methods research design, both quantitative and qualitative data were collected in this study, through a background questionnaire, holistic scores given to translated texts and TAPs. The data obtained from the rater profile forms that the raters were asked to fill in before the scoring task were used to categorize them according to their characteristics for the interpretation of both the quantitative and the qualitative data sets. The quantitative data of the study involved 80 translation scores assigned to 80 translated texts by each of the raters who used a holistic scoring scale (see Appendix D) developed by Marais (2013). The qualitative

data collected from the TAPs that the raters recorded at the time of the rating process were used to explore the raters' decision-making behaviors. In the following sections, detailed information is provided about the instruments used for data collection.

Selection of Translation Passages

The texts used in this study were selected in accordance with the following steps:

- Texts were selected from authentic sources, rather than the ones prepared for educational purposes.
- Texts were related to daily life issues.
- Texts were easy to read and suitable for the language proficiency level of the students.
- The length and difficulty of the texts were determined according to their correspondence to the samples that students translate in their translation classes.
- Texts did not include any terminological difficulties.

In this study, four informative newspaper articles obtained from an online British magazine (Daily Star Online) were selected to be translated by the students in four different sessions controlled by the researcher. Four different informative articles from a newspaper (see Appendix E) were chosen since they can be categorized as authentic texts. As the act of translation is carried out purely for communicative purposes in the real world, students should also be given contextualized and real life-related tasks (Carreres & Noriega-Sánchez, 2011). For this reason, studies carried out on translation specifically make use of authentic texts such as news summaries (Conde, 2011), advertising articles (Eyckmans, Anckaert & Segers, 2009), press articles (Carreres & Noriega-Sánchez, 2011), film subtitles (Carreres & Noriega-Sánchez, 2011) and literary texts (Shahraki & Karimnia, 2011).

In each translation session, all 40 students participating in the study were asked to translate a given newspaper article (4 in total), from English to Turkish. The students were given 90 minutes to complete each translation task and the translations were completed with pen and paper. The students were allowed to use any type of dictionary (manual or electronic) they wanted while translating. However, the students were not allowed to interact with each other during the translation sessions and any necessary explanations were made in the mother tongue (in Turkish). The whole process was controlled by the researcher.

Overall, 160 translation papers were obtained from the students who participated in the study. Before the collection of the main data from the raters, the researcher carefully divided the collected translation papers regarding their quality (high and low) with another expert rater

by following the criteria included in the holistic rubric used in the study. At the end of this procedure, (40) high-quality and (40) low-quality translation papers were randomly selected to be used in the study.

Readability of the articles.

Before administering the translation tasks to the students, the researcher conducted a test (The Flesch-Kincaid Test) on each text in order to determine its degree of readability. The Flesch-Kincaid test is one of the procedures widely used in determination of the difficulty of a text (Kincaid, Fishburne Jr., Rogers & Chissom, 1975). In the Flesch Reading Ease test, the total number of syllables, words, and sentences in the text is considered and the score ranges from 0 to 100. The closer the score to 0, the more difficult the text is. For each article used in the study, the researcher conducted a readability test to determine the Flesch Reading Ease score using the Readability Analyzer program. The results showed that the texts chosen by the researcher had scores between 62.5 and 69.96 for Flesch Reading Ease. This means that the texts had a grade eight level of difficulty and could be read by the average student in the specified grade level. As the texts were neither too easy nor too difficult, they were found to be suitable for college students. Therefore, the researcher considered the texts to be ideal for the research project. Table 6 shows the statistics for the passages obtained from Readability Analyzer program.

Table 6. *Readability Analyzer Statistics of the Four Passages*

	Passage 1	Passage 2	Passage 3	Passage 4
Number of Words	301	338	300	322
Number of Sentences	18	18	17	17
Percentage of Difficult Words	12.05%	9.12%	12.06%	8.17%
Flesch Reading Ease	62.52	66.4	62.5	69.96
Flesch-Kincaid Grade Level	8.16	8.54	8.66	8

Characteristic features of the articles.

All news articles reported a specific incident (daily events) or development that happened recently. The register of the articles was formal, objective and no personal opinions were included. Each of the articles had an eye-catching headline stating the topic of the news in a short, strong statement. The first leading paragraph of the articles consisted of specific information about the '5 W's' (who, what, when, where and why). The supportive body paragraphs included information about how the incidents took place with detailed descriptions. The sentences in each paragraph were typically straight-forward and short. Mostly, simple declarative sentences, composed of common, everyday words were used to convey ideas. All

the paragraphs were brief, including a few sentences. The number of words in these articles was between 300 and 350, which is more or less similar to the number of words in Kim's model text (2009).

Rating Scale (Holistic Rubric) Used in the Study

A rubric can be defined as an evaluation instrument that categorizes expected learning outcomes and divides them into different performance levels (Sezer, 2005). Rubrics are used in language testing and evaluation for measuring many features and competences that must be acquired in language production (Cohen 1994). A rubric usually contains all the sub-components that constitute a structure and is developed by defining what is evaluated (e.g., translation skill) (Angelelli & Jacobson, 2009, p.38). In performance evaluation, two types of rubrics are used: holistic and analytic. While holistic rubrics focus on the product or process as a whole and allow for a very general evaluation without disrupting performance, analytic rubrics divide the product or process into various dimensions, allowing each skill involved in performance to be evaluated independently (Birel & Albuz, 2014).

The advantages of using rubrics in evaluation are as follows. They serve as a reference for the raters that facilitates the decision-making process within existing transparent criteria by limiting the subjectivity inherent in any evaluation process that includes human intervention. In order to provide a general evaluation, it is adequate to give only positive or negative values to each of the descriptors. The use of rubrics provides a general evaluation of the text with a holistic approach in the assessment of a translation as a product. As performance sub-dimensions are predetermined, they facilitate the task of the evaluator (Jiménez-Crespo, 2011, p. 76). Unlike traditional evaluation techniques, rubrics provide students with more informative feedback about their strengths and the skills that need to be improved (Andrade, 2000). It can be said that the use of rubrics makes the provision of reliability in measurement easier and feasible and helps teachers transfer the evaluation criteria to the students (Sezer, 2005, p. 68). Rubrics explain what is expected of students, facilitate feedback and allow self-assessment and peer assessment. In addition, in order for an evaluation list to fulfill its purpose, it must create different performance levels for each direction to be evaluated and specify a rating (weak / good / perfect; weak / insufficient / sufficient; 0 to 5 or 10; etc.) (Hurtado Albir, 2015).

The rating scale used in this study was developed by Marais (2013) for the purpose of evaluating student translation products in educational settings. The rubric consists of six sections structured hierarchically: 1) suitability of the translation for the general purpose, 2) culture and target reader, 3) text, 4) design, 5) content and 6) language assessment. In this hierarchy, the sections are not ranked in order of importance and the main objective is to reflect

the decisions taken by the students during their translation process. The structure of the rubric is based on a holistic approach. Performance scores have been assigned to the subcategories of the rubric sections: 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 highest score a student can get from a translation paper is determined as 10 in this rating scale. According to the developer of this rubric (Marais, 2013), a translated product should be good in all aspects and as a whole. Thus, the requirements for a student to pass a translation assignment are 50%, i.e. 5 or more out of 10. Moreover, what is emphasized here is that if any section of a translation fails, then the whole translation fails. For this reason, in application of this translation assessment instrument, a student should get at least half of the point assigned for each of the sections. Otherwise, the translation is marked as a poor/weak one.

It is essential that the assessment scales are shared with students before the relevant tasks are fulfilled, so that students can better grasp what is expected of them and become aware of the criteria for evaluation (Hurtado Albir, 2015; Khanmohammad & Osanloo, 2009). Therefore, the holistic rubric and a checklist used in this study were shared with the students before the application of the translation tasks. Figure 5 shows the score weight distribution of the six components of the holistic scale used in the study.

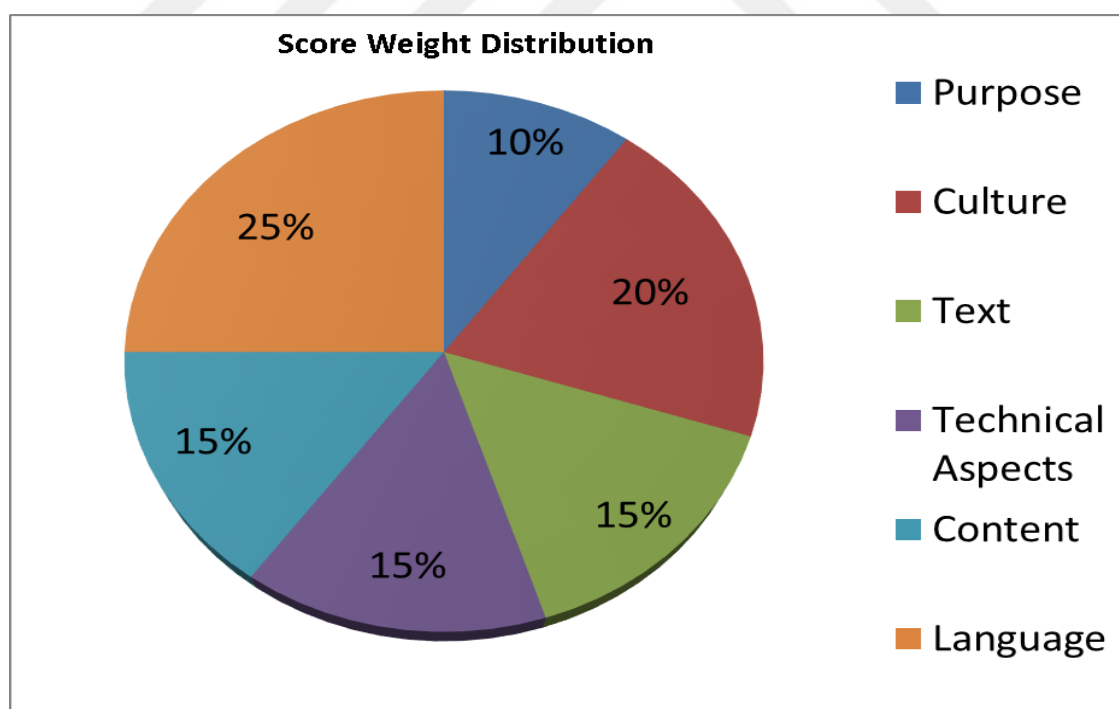


Figure 5. Distribution of the score weights to the rubric's subscales.

Think-Aloud Protocols (TAPs)

According to Charney (1984), reliable scores can be achieved through superficial and quick rating rather than following an in-depth analysis of the text. On the other hand, some

researchers claim that evaluations including scoring rubrics and that are based on deeper interpretations of the texts can yield more reliable and valid scores (Cumming, Kantor & Powers, 2001; Huot, 1993). In this sense, many studies conducted in the area of second/foreign language writing assessment (e.g., Barkaoui, 2010; Connor-Linton, 1995; Gebril & Plakans, 2014; Han, 2017) have applied TAPs as an efficient way of determining the decision-making processes of the raters. By including TAPs in this study, the researcher aimed to examine the internal decision-making processes of the raters while evaluating the translation papers. Thinking aloud can be described as a technique that enables a subject to think loudly and express whatever goes through his mind while performing the task in hand (Van Someran, Bernard & Sandberg, 1994, p. 26). Wolfe, Kao and Ranney (1998) identify this method as a cognitive process that is composed of various mental states, each representing the final product of the processed information. Figure 6 depicts how think aloud memory works.

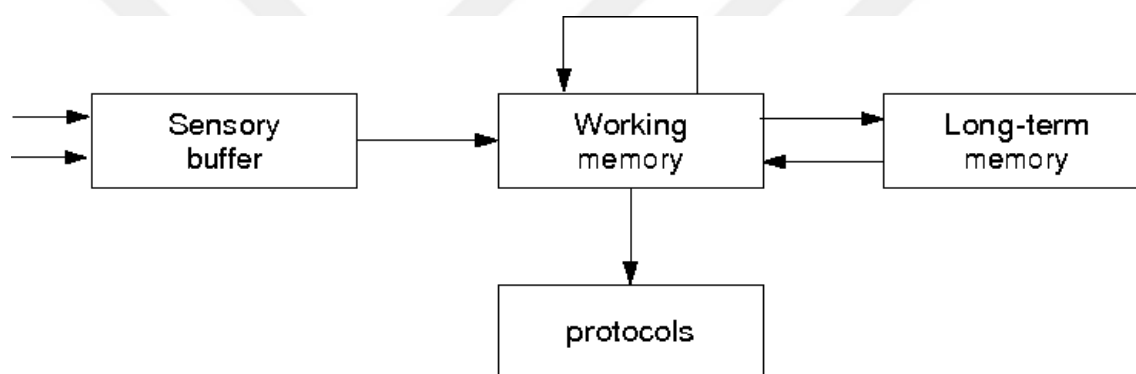


Figure 6. The think aloud memory.

In contrast to the other techniques applied for collecting verbal data, TAPs allow the raters to concentrate on the task in hand and encourage them to express their exact thoughts by allowing them to use their own language without any interruptions (Van Someran *et al.*, 1994, p. 26). This verbal reporting method includes five cognitive processes: perception, retrieval, construction, storage and verbalization. In the last process, the general information that already exists in the long-term memory is verbalized, and this procedure may be performed by applying techniques such as retrospection, introspection, questions and prompting, and dialogue observation (Van Someran *et al.*, 1994). Table 7 shows the techniques related to the think aloud method.

Table 7. *Techniques Related to Think Aloud Method*

	Disturbance	Memory errors	Interpretation
Retrospection:	no	yes	yes
Introspection:	no	little	yes
Prompting:	yes	no	little
Dialogues:	not applicable	no	no

Combining introspection technique with TAPs.

Introspection is used to obtain a transient model of the structure of the cognitive process. Then, the method of thinking aloud is used to validate this model with more direct data. As applied by psychologists in the 1920s and 1930s, the subject is asked to give a correct, complete and comprehensible report on a cognitive process (Van Someren *et al.*, 1994, s. 23). In the introspection technique, the raters verbalize their thoughts while completing the task, but at interval points determined by them. This simple verbalization process provides an objective method as the data obtained from verbal protocols are accessible to everyone.

Studies including think aloud protocols especially emphasize that affective factors have a significant impact on translation assessment (Laukkanen, 1996). For this reason, the use of TAPs is essential in understanding and identifying the factors that affect the scoring process of the translation products. In this study, the raters were asked to use TAPs with a voice-recording device while scoring 32 pre-determined (randomly) translation papers (16 low-quality and 16 high-quality) that were in their translation paper pack. The raters were asked to provide an accurate, clear and consistent report of this cognitive process by application of the introspection method in the evaluation process. These records were then transcribed and the raters' evaluation based utterances were analyzed. In the analysis of these data, the coding and categorization were done by the researcher and the effective factors which were likely to be included in the character and origin of the performance differences between the raters were determined.

Before collection of the main data from the raters, the researcher scored one of the translation papers in order to show the raters how to conduct TAPs, and after the scoring process was video-recorded, it was uploaded to YouTube (Aybirdi, 2019). Then, it was shared with the raters. Also, in order to make sure that the raters understood the think aloud procedure fully, the researcher conducted a TAP training process. Before the collection of the TAPs data from the raters, a training session was held by the researcher about the application of this method along with the holistic rubric. First of all, the reason why the method of TAP was included in the research was explained to the raters and then some instructions were given related to the use of the method. Lastly, the raters were asked to make suggestions about how to conduct TAPs efficiently.

Data Collection Procedure

After the administration of the four translation tasks to the students, prepared data packs, including a copy of the original four texts translated by the students, 80 translated texts, 80 holistic scoring rubrics, a checklist for the rubric and a background information questionnaire

were handed to the 10 raters. Also, voice recorders were provided to the raters who claimed not to have a suitable device for voice-recording. The raters were given a two-month period of time (from the beginning of April to the end of May 2019) to score the translated papers.

Both quantitative and qualitative data collection procedures were followed in this study. While the quantitative data set consisted of a total number of 5,600 scores (800 total scores and 4,800 sub-scores), the qualitative data set was composed of 256 TAPs that the raters voice-recorded while scoring the pre-determined translated texts. All of the raters were required to use TAPs while scoring the papers. However, in spite of the sample video and training provided to the evaluators, 2 of them failed to complete the thinking aloud tasks as required by the research. Regarding the language preferences of the raters while recording TAPs, while 3 of the raters used English, the rest of them used Turkish. Also, the researcher made connection with the raters at intervals in order to handle possible problems that might stem from the TAPs. The following sections provide further information about the data collection procedures used in this study.

Rating Procedure

In this study, each of the raters used a 10-point holistic rubric in order to score 80 translated texts regarding six aspects: purpose, culture, text, design, content and language (800 total scores and 4.800 sub-scores). The raters were free to use partial points while scoring each sub-category within the given score bands in the rubric (e.g. purpose=1.0, culture=0.5, text=1.3, design=0.8, content=1.2, language=1.0, and total score=5.8). The raters were asked to follow some evaluation guidelines (Appendix F) while recording TAPs for predetermined translation papers in order to ensure standardized evaluation process. Also, a checklist created by the developer of the holistic rubric used in this study was provided to the raters for convenience in commenting. In order to avoid unfair judgment, the raters were not allowed to negotiate information with each other regarding the translation papers. Also, the raters were allowed to take notes on the papers. In order to create a more authentic rater reaction to the translation papers, the raters were told to feel comfortable in giving feedback as if the papers were the products of their own students. A two-months time limit was set for the raters to score the papers since they should not feel under pressure while scoring each paper. After completing the scoring procedure, the raters were required to return the data packs to the researcher.

Voice-recording Raters' Verbal Thoughts

The raters who participated in this study were trained to record TAPs in which they verbalized their thoughts while scoring students' translation papers. The raters were asked to

keep separate voice-recordings for each of the pre-determined translated papers (16 high-quality and 16 low-quality translation papers). The reason for this was to prevent raters from commenting on previously scored papers accidentally. Otherwise, such confusion would cause extreme data loss.

From beginning to the end of the scoring process, the raters were required to continue verbalizing their thoughts. Thus, the whole grading procedure would be examined. Also, they were reminded that they were not required to rationalize their thoughts at length, since the main aim of the study was to determine their natural assessment processes. Regarding the language to be used while recording TAPs, the decision about whether to use English or Turkish was left to the discretion of the raters. At the end of the assessment procedure, the raters were asked to deliver the recordings, translation papers and rubrics to the researcher.

Role of the Researcher

In this study, the researcher took part in several phases; Firstly, the researcher controlled the process of administering the four translation tasks. The administration of these tasks was carried out under real-life examination conditions. Secondly, before the collection of the main data from the raters, the researcher divided the collected translation papers regarding their quality (high and low) with another expert rater. Thirdly, the researcher transcribed the qualitative data obtained from the TAPs, and then coded the transcribed data according to a coding scheme. To minimize the errors from coding and to increase the reliability of the data, another expert coder was employed to work on the same coding task.

Validity and Reliability

Validity of the holistic rubric.

The rubric used in the study was developed for use in pedagogical contexts. However, the rubric was submitted for approval to a group of experts in the field of English Language Education and Educational Psychology to check its suitability and applicability. The internal validity of the evaluation rubric was determined as 0.895 by calculating the internal consistency of its different parts.

Single rater reliability.

Each example of the scoring performed by a rater whose scoring was ideally based on fixed and objective criteria was compared with other scoring examples performed by the same rater.

Inter-rater reliability.

When the scoring raters received training on the structure and use of the reference (e.g., specific norms and criteria, procedures, etc.), inter-rater reliability increased as there was more differentiation in all dimensions. According to Lievens, in scoring, the evaluation can be more accurate and consistent when a specific rating scale is used (2001, pp. 255-257). Since the exchange of ideas between the raters weakens the consistent and objective quality of scoring by threatening the reliability of the scoring, the most important factor regarding reliability is that each scoring must be made completely independently of the others in order to maintain the integrity of the scoring criteria and the scoring procedure (Angelelli, 2009, p. 19). Therefore, the raters who participated in this study were asked not to contact other raters during the scoring process.

Preparation of the Data

The Microsoft Excel program was used in the preparation of both the qualitative and the quantitative data sets for analysis. Also, some other computer programs were applied during the analysis such as SPSS Statistics program 25.0 (for inferential and descriptive statistics of both data sets) and EduG software program 6.0 (for generalizability analysis of the quantitative data set).

Preparation of quantitative data.

The scores that the raters gave to translation papers composed the quantitative data set. A total number of 5,600 translation scores (800 total scores and 4,800 sub-scores) obtained from 80 translation papers scored by 10 raters were processed using the SPSS program.

Descriptive and inferential statistics.

Descriptive and inferential analyses were performed both on the total scores given to the translation papers and on the sub-scores assigned to the six components of the papers. Also, descriptive statistics analysis was performed on the codes obtained from the TAPs' analysis.

G-theory analyses.

In order to understand whether students and raters and the interactions between them contribute to the variance in the scores assigned to the translation papers, the G-theory framework was applied in this study using the EduG software program. Also, calculations of generalizability and dependability coefficients were performed in order to understand whether the reliability of the scores given to low- and high-quality translation papers showed difference

in relation to the raters' experience of rating. In this study, while the papers produced by the students were the object of measurement, raters and translation quality were random facets. As the papers (papers as p) were produced by the students, and the raters (r) assigned scores to the high- and low-quality translation papers (q), the G-study design was completely crossed as $p \times r \times q$. The G-studies employed in this study were as follows;

- 1) Paper-by-rater-by-quality ($p \times r \times q$) random effect G-study was employed in order to determine independent sources of variation such as persons (p), raters (r), quality (q), paper-by-rater ($p \times r$), paper-by-quality ($p \times q$), rater-by-quality ($r \times q$), paper-by-rater-by-quality ($p \times r \times q$) 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, paper-by-rater ($p \times r$) random effects G-study was employed; papers (p), rater (r), paper-by-rater ($p \times r$) for 40 low-quality translation papers scored by holistic evaluation methods.

Also, calculations of generalizability and dependability coefficients were carried out to determine the reliability of the data set.

- 3) In order to attain variance component estimates, paper-by-rater ($p \times r$) random effects G-study was employed; papers (p), rater (r), paper-by-rater ($p \times r$) for 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.

Furthermore, since the raters were categorized into two groups in respect of their previous rating experience as experienced and inexperienced, the scores they assigned to the translation papers were compared in terms of their generalizability and dependability coefficients. Thus, paper-by-rater-by-quality ($p \times r \times q$) random effects G-study was performed on all translation papers, and paper-by-experience ($p \times r$) random effects G-study was conducted on high- and low-quality translation papers. Also, the same procedure was followed to determine differences between female and male raters.

Preparation of qualitative data.

The qualitative data set was composed of the TAPs that the raters recorded for 32 of the translation papers while evaluating them. The papers that were to be voice recorded were

randomly selected by the researcher and “TAP” was written with a highlighter pen at the top of these papers for the attention of the raters. While analyzing the data obtained from the TAPs, content analysis method was employed. In doing so, a top-down (deductive) approach was followed. The analysis of the data was carried out by application of a coding scheme adapted from Cumming, Kantor and Powers (2002). The phases followed in this process are as follows:

1. The collected data were transcribed by the researcher.
2. The data were divided into meaningful segments.
3. The coding frame was discussed with two experts in the field.
4. A piloting coding was applied for the coding scheme.
5. The coding scheme was evaluated and notified with a field expert.
6. For inter-rater reliability, an independent researcher code 15% was attained.
7. The coding scheme was reevaluated with two experts in the field.
8. The data were coded.

The total duration of the TAPs was 36 hours and 8 minutes. The length of the evaluation time for each translation paper was approximately 7:20 minutes. The data obtained from these protocols were divided into meaningful units by applying a coding scheme. The segmented units were composed of some single words and a variety of sentences related to the same aspect of the evaluated papers. In this study, the raters assigned scores to the translation papers one at a time, thus they recorded separate voice files for each of the papers. 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 the qualitative data analysis was asked to code a randomly chosen sample of 15% of the protocols (Lumley, 2002). The analysis carried out to determine the similarity between the coding procedures indicated an agreement of .80. After inter-rater reliability between the coders was achieved, a follow-up discussion session was held in order to detect the coding differences between the two researchers. After the piloting and inter-rater reliability procedures

had been completed, various changes were made in the coding frame in consultation with two other experts in the field.

After the last revisions of the coding scheme, another discussion session was held regarding the coding of the TAPs. The final agreement reached by the two coders was determined as $k = .82$.

Research Ethics

In this study, there are no anticipated risks, and the requirements of research ethics were fulfilled by the researcher during the procedures of data collection and thesis writing. First of all, the students who participated in the study to fulfill the translation tasks were provided with necessary information regarding the purpose of the research. Also, they signed a written consent form stating that they were aware of their rights and the potential risks of the research before the administration of the tasks. Secondly, the necessary official permissions were obtained from the administration of Kafkas University Faculty of Science and Letters where the students are enrolled (see Appendix G). Thirdly, the 10 raters who participated in the research were informed about the background, purpose, risks, benefits, significance, and procedures of this research. Also, the raters were provided with a consent form assuring confidentiality of their identities and stating their rights (see Appendix H). Fourthly, for the quality categorization of the translation papers, the researcher contacted two other experts in the field and each of them signed the consent form. Therefore, all of the students, raters and experts who took part in this study were assured that participation was voluntary, that their identities would be kept confidential and that they had the right to withdraw from the study whenever they might want to without giving any excuse.

Summary

In this chapter, the methodological aspects of the present study were explained in detail. First of all, all theoretical frameworks (Dynamic Equivalence Theory, Constructive Alignment and Generalizability Theory) guiding the study were described and a description of the demographics of the students and raters who participated in the study was provided in detail. Following the presentation of the data collection instruments, detailed information was given regarding the procedures of the quantitative and qualitative data collection. Besides, the steps that were followed in the data preparation process and the computer programs used in analysis of the data were presented. The results and findings obtained from the study were provided in the next chapter.

CHAPTER FOUR

Findings

In this chapter, the findings pertaining to each question of the research are presented individually. Since this study was conducted with a convergent parallel design, the results obtained from the qualitative and quantitative data sets are analyzed and presented separately. Firstly, the results of the quantitative data analysis are presented to respond to the first four research questions. Afterward, the results of the qualitative data analysis are given to respond to the last research question.

CFA Analysis Results for Validity and Reliability of Rubric

Confirmatory Factor Analysis (CFA) results for the rubric were given in Figure 7.

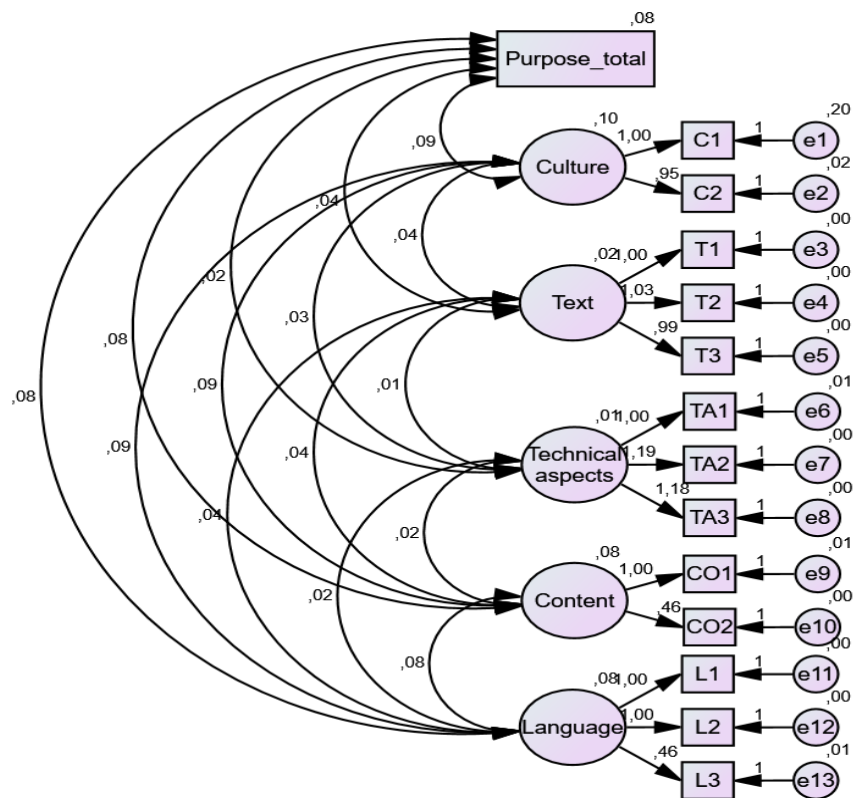


Figure 7. CFA results for rubric.

Table 8. CFA Results and Fitness Criteria

Criteria	Good fit	Acceptable fit	Results
$\bar{\chi}^2$	$0 \leq \bar{\chi}^2 \leq 2df$	$2df < \bar{\chi}^2 \leq 3df$	269.265
P	$.05 < p \leq 1.00$	$.01 \leq p \leq .05$	0.000
$\bar{\chi}^2/df$	$0 \leq \bar{\chi}^2/df \leq 2$	$2 < \bar{\chi}^2/df \leq 5$	4.274
RMSEA	$0 \leq RMSEA \leq .05$	$.05 < RMSEA \leq .08$	0.064
P	$.10 \leq p \leq 1.00$	$.05 \leq p \leq 1.00$	0.000
NFI	$.95 \leq NFI \leq 1.00$	$.90 \leq NFI < .95$	0.985
CFI	$.97 \leq CFI \leq 1.00$	$.95 \leq CFI < .97$	0.988
GFI	$.95 \leq GFI \leq 1.00$	$.90 \leq GFI < .95$	0.955
AGFI	$.90 \leq AGFI \leq 1.00$	$.85 \leq AGFI < .90$	0.925

GFI(Goodness of Fit Index), AGFI(Adjusted Goodness of Fit Index) NFI (Normed Fit Index), CFI(Comparative Fit Index), RMSEA(Root Mean Square Error of Approximation).

Source: Schermelleh & Moosbrugger, 2003

The CFA analysis results and criteria showed that the holistic rubric selected for the scoring procedure and raters' evaluations had a good and acceptable fit ($p < 0.001$). Thus, it was proven that the results obtained from the use of holistic rubric were acceptable for the analysis of the research questions.

For reliability analysis, the Cronbach Alpha level of the rubric was evaluated. Table 9 shows results of the reliability analysis.

Table 9. Reliability Analysis Results

Cronbach's Alpha	N of Items
.895	21

The reliability analysis results showed that the scores assigned to the rubric as a whole and to its sub-categories had internal consistency. In the literature, an alpha level above 0.60 is acceptable, and an alpha level over 0.90 indicates redundancy in the rubric. Thus, the 0.895 alpha level showed that the rubric had acceptable reliability with no redundancy.

Quantitative Data Analysis Results

The SPSS 24.0 program was used to generate descriptive and inferential statistics to answer the first two research questions. While the first research question asked whether there were any significant differences between the holistic scores assigned to high- and low-quality translation papers, the second research question inquired whether there was variability between and among raters' scores given to each component and sub-component of the rubric. G-theory analysis was performed to address the third research question, which attempted to determine sources of score variation contributing to the variability in holistic scores; and the fourth

research question, which inquired whether the reliability of holistic scores assigned to high- and low-quality translation papers showed difference based on the raters' gender and previous rating experience.

Findings Related to Research Questions

The five research questions guiding this thesis study were evaluated as given below:

1. Are there any significant differences among the holistic scores assigned to low- and high-quality translations?
2. What is the score variability between and among raters' scores assigned to each component and sub-component of the rubric?
3. What are the sources of score variation that contribute to the score variability in the holistic scores assigned to high- and low-quality translation papers?
4. Do the variables of gender and rating experience have an impact on the reliability of the scores assigned to high- and low-quality translation papers?
5. How do raters make decisions while rating different quality translations holistically?

Results for the first research question.

The first research question is given below:

“Are there any significant differences among the holistic scores assigned to low- and high-quality translations?”

The Kolmogorov Smirnov Test was used for normality of scores. Since all scores were distributed non-normally, nonparametric tests were used for the analysis. The Mann-Whitney U test results for differences between low- and high-quality translations are given in Table 10.

Table 10. *Mann-Whitney U test Results for Differences Between Low- and High-Quality Translations*

	U	p	High Quality (Mdn)	Low Quality (Mdn)
Purpose 1	9383.000	0.000	0.800	0.400
Purpose total	9383.000	0.000	0.800	0.400
Culture 1	8097.000	0.000	0.800	0.300
Culture 2	7778.500	0.000	0.800	0.300
Culture total	7628.000	0.000	1.600	0.600
Text 1	11563.000	0.000	0.400	0.200
Text 2	10052.000	0.000	0.400	0.200
Text 3	10997.500	0.000	0.400	0.200
Text total	8624.500	0.000	1.200	0.600
Technical aspects 1	35269.000	0.000	0.400	0.200
Technical aspects 2	24897.500	0.000	0.400	0.200

Table 10. (continued)

Technical aspects 3	28478.500	0.000	0.400	0.200
Technical aspects total	24614.500	0.000	1.200	0.700
Content 1	10064.500	0.000	0.800	0.300
Content 2	11315.500	0.000	0.400	0.200
Content total	7767.500	0.000	1.300	0.500
Lang 1	9127.500	0.000	0.800	0.300
Lang 2	9150.000	0.000	0.800	0.300
Lang 3	11810.500	0.000	0.400	0.200
Lang total	7673.000	0.000	2.000	0.800
Grand total	6875.000	0.000	8.400	3.800

According to Table 10, the total purpose score median was 0.8 in high quality texts, and 0.4 in low quality texts with statistically significant difference ($U=9383.000$; $p<0.05$). The purpose scores of high-quality texts were significantly higher than those of low quality texts.

The culture score median was 1.6 in high quality texts, and 0.6 in low quality texts with statistically significant difference ($U=7628.000$; $p<0.05$). The culture scores of high-quality texts were significantly higher than those of low quality texts. Culture had two sub-scores, and the medians of these two items were similar in both within groups. Thus, the differences were also significant for both items between groups ($p<0.05$).

The text score median was 1.2 in high quality texts, and 0.6 in low quality texts with statistically significant difference ($U=8624.500$; $p<0.05$). The text scores of high-quality translations were significantly higher than those of low-quality translations. The text score had three sub-scores and their medians were the same for both within group comparisons, and were significantly different between groups ($p<0.05$).

The technical aspects score median was 1.2 in high quality texts, and 0.7 in low quality texts with statistically significant difference ($U=24614.500$; $p<0.05$). The technical aspects scores of high-quality texts were significantly higher than those of low quality texts. The technical aspects score had three sub scores and their medians were the same for both within groups comparisons, and were significantly different between groups ($p<0.05$).

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 content scores of high-quality texts were significantly higher than those of low-quality texts. The content score had two sub-scores and their medians were not the same for both within groups comparisons, and were significantly different between groups ($p<0.05$).

The language score median was 2.0 in high-quality texts, and 0.8 in low-quality texts with statistically significant difference ($U=7673.000$; $p<0.05$). The language scores of high-quality texts were significantly higher than those of low-quality texts. Language score had three sub-scores and their medians were not the same for both within groups comparisons, and were significantly different between groups ($p<0.05$).

Overall, the median for grand total for high-quality texts was 8.4 and higher than the low-quality text median (3.8), with a statistically significant difference ($U=6875.000$; $p<0.05$). Schemes 4.2 and 4.3 show the median values' deviation of and the ranges for scores given to high- and low-quality translation papers. Since they split the data into quartiles, boxplots are useful for visualizing the data distribution. The body of the boxplot displays 50% of the data. Representing the second quartile, the horizontal black line gives the median of the data set. On the other hand, the vertical lines in the box are whiskers that show the remaining 50% of the data. The data from the smallest non-outlier to the first quartile are shown through the line that extends below the box. The whisker above the box displays points of data from the third quartile to the largest non-outlier. While the skewness patterns of the data are determined through the length of the quartiles, the range is determined through the vertical distance of the boxplot. The distribution of grand total scores for high-quality translation passages was given in Figure 8.

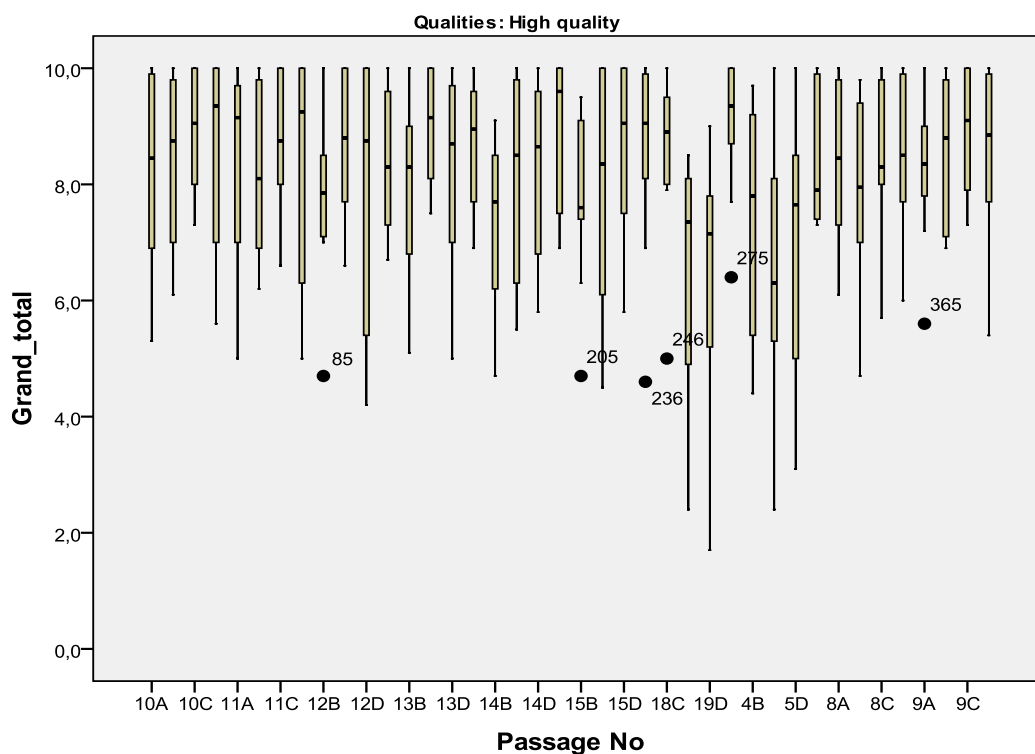


Figure 8. Distribution of grand total scores for high quality translation passages.

According to Figure 8, papers 19D, 18C and 5D had greater ranges than the other high-quality translation papers. However, all paper medians and means were close to each other. The distribution of the grand total scores for low-quality translation passages is given in Figure 9.

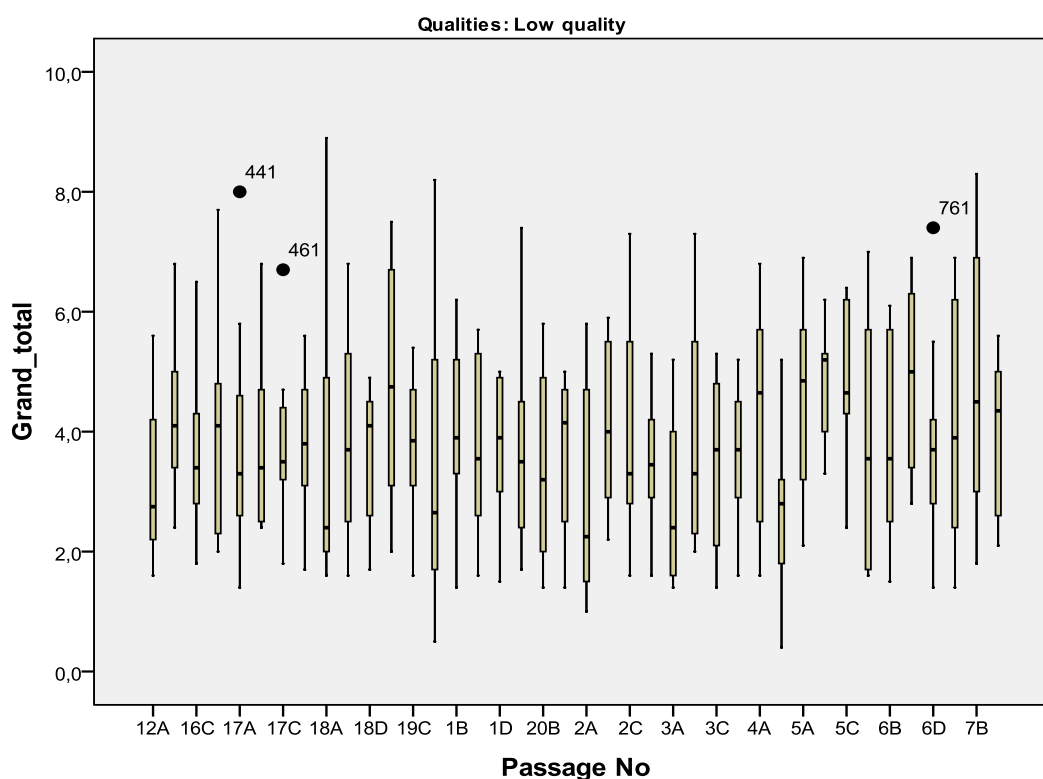


Figure 9. Distribution of grand total scores for low quality translation passages.

In low quality translation papers, the median and means were close to each other, but had greater ranges than did the high quality texts. Thus, it might be argued that there was more variation in the evaluation of low-quality texts than in that of the high quality translation papers. In other words, the evaluations of high-quality translation papers were more similar than the evaluations of low-quality papers. When compared to the high-quality translation papers, there appears to be more variance in the median scores of the low-quality translation papers. This suggests that raters were more consistent when assigning scores to high-quality papers than to the low-quality ones.

Results for the second research question.

The second research question is given below:

“What is the score variability between and among the raters’ scores assigned to each component and sub-component of the rubric?”

A Kruskal Wallis test was performed to determine variability between and among the raters’ scores given to each component and sub-component of the holistic scale. The Kruskal

Wallis test results for score variability between and among the raters' scores given to each component of the scale are given in Table 11.

Table 11. *Kruskal Wallis Test Results for Score Variability Between and Among Raters' Scores (Mdn)*

	X²	p	1	2	3	4	5	6	7	8	9	10
Purpose total	74.045	0.000	0.60	0.80	0.60	0.50	0.40	0.70	0.70	0.65	0.50	0.60
Culture total	65.984	0.000	1.20	1.15	1.30	0.80	0.65	1.20	1.40	1.20	1.00	1.20
Text total	55.266	0.000	1.20	1.00	0.95	0.60	0.75	0.90	1.20	1.00	0.95	0.90
Technical aspects total	145.970	0.000	1.00	1.00	1.40	0.75	0.70	0.90	1.20	1.20	0.90	1.15
Content total	63.260	0.000	1.00	1.00	1.15	0.55	0.60	1.00	1.10	1.15	0.95	0.95
Lang total	61.078	0.000	1.60	1.70	1.55	0.85	1.00	1.45	1.70	1.45	1.55	1.50
Grand total	68.249	0.000	6.35	6.70	6.55	4.10	4.15	6.30	7.00	6.35	5.90	6.45

As Table 11 illustrates, for each category of the rubric, the total scores showed statistically significant differences ($p < 0.05$). For the purpose category, the 2nd rater evaluation was the highest, and the 5th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 74.045$; $p < 0.05$). For the culture category, the 7th rater evaluation was the highest, and the 5th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 65.984$; $p < 0.05$). For the text category, the 1st and the 7th rater evaluations were the highest and the 4th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 55.266$; $p < 0.05$). For the technical aspects category, the 3rd rater evaluation was the highest and the 5th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 145.970$; $p < 0.05$). For the content category, the 3rd and the 8th rater evaluations were the highest, and the 5th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 63.260$; $p < 0.05$). For the language category, the 2nd and the 7th rater evaluations were the highest, and the 4th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 61.078$; $p < 0.05$). In among the grand total scores, the 7th rater's rating was the highest, and the 4th rater evaluation was the lowest of the rater scores with significant differences ($X^2 = 68.249$; $p < 0.05$). In general, it was found that there was not a linear or regular evaluation trend among raters, and that the raters' scores varied between and among raters for each category of the rubric.

The Kruskal Wallis test results for score variability between and among raters' scores given to each sub-component of the scale are given in Table 12.

Table 12. *Kruskal Wallis Test Results for Score Variability Between and Among Raters' Scores (Mdn)*

	X²	p	1	2	3	4	5	6	7	8	9	10
Purpose 1	74.045	0.000	0.60	0.80	0.60	0.50	0.40	0.70	0.70	0.65	0.50	0.60
Purpose total	74.045	0.000	0.60	0.80	0.60	0.50	0.40	0.70	0.70	0.65	0.50	0.60
Culture 1	68.222	0.000	0.60	0.60	0.65	0.40	0.30	0.65	0.70	0.60	0.50	0.60
Culture 2	62.007	0.000	0.60	0.60	0.65	0.40	0.30	0.60	0.70	0.60	0.50	0.60
Culture total	65.984	0.000	1.20	1.15	1.30	0.80	0.65	1.20	1.40	1.20	1.00	1.20
Text 1	63.479	0.000	0.40	0.30	0.35	0.20	0.25	0.30	0.40	0.35	0.35	0.30
Text 2	59.306	0.000	0.40	0.30	0.30	0.20	0.20	0.30	0.40	0.30	0.30	0.30
Text 3	48.496	0.000	0.40	0.30	0.30	0.20	0.30	0.30	0.40	0.35	0.30	0.30
Text total	55.266	0.000	1.20	1.00	0.95	0.60	0.75	0.90	1.20	1.00	0.95	0.90
Technical aspects 1	157.596	0.000	0.35	0.30	0.50	0.20	0.20	0.30	0.40	0.40	0.35	0.30
Technical aspects 2	108.189	0.000	0.35	0.30	0.45	0.30	0.20	0.30	0.40	0.40	0.30	0.45
Technical aspects 3	123.992	0.000	0.30	0.30	0.50	0.20	0.20	0.30	0.40	0.40	0.30	0.40
Technical aspects total	145.970	0.000	1.00	1.00	1.40	0.75	0.70	0.90	1.20	1.20	0.90	1.15
Content 1	75.649	0.000	0.60	0.70	0.75	0.35	0.35	0.65	0.75	0.75	0.55	0.60
Content 2	59.156	0.000	0.40	0.30	0.35	0.20	0.30	0.30	0.40	0.40	0.30	0.30
Content total	63.260	0.000	1.00	1.00	1.15	0.55	0.60	1.00	1.10	1.15	0.95	0.95
Lang 1	58.888	0.000	0.60	0.70	0.55	0.35	0.40	0.60	0.60	0.50	0.60	0.60
Lang 2	74.366	0.000	0.60	0.70	0.50	0.30	0.30	0.60	0.70	0.60	0.65	0.50
Lang 3	40.027	0.000	0.40	0.30	0.35	0.25	0.30	0.30	0.30	0.30	0.35	0.30
Lang total	61.078	0.000	1.60	1.70	1.55	0.85	1.00	1.45	1.70	1.45	1.55	1.50
Grand total	68.249	0.000	6.35	6.70	6.55	4.10	4.15	6.30	7.00	6.35	5.90	6.45

Table 12 shows that the first purpose category median and the total purpose category mean was the highest for the 2nd rater, and the lowest for the 5th rater and that the difference between them was statistically significant ($X^2=74.045$; $p<0.05$). The first culture sub-category median was the highest for the 7th rater, and the lowest for the 5th rater with statistically significant difference ($X^2=68.222$; $p<0.05$). The second culture sub-category median was highest for the 7th rater, and the lowest for the 5th rater with statistically significant difference ($X^2=62.007$; $p<0.05$). The total culture category median was the highest for 7th rater, and the lowest for 5th rater with statistically significant difference ($X^2=68.222$; $p<0.05$). Thus, it can be seen from Table 12, the culture category scores vary between raters, but do not vary among raters.

The first text sub-category medians were highest for the 1st and the 7th raters, and lowest for the 4th rater with statistically significant difference ($X^2=63.479$; $p<0.05$). The second text sub-category medians were highest for the 1st and the 7th raters, and lowest for the 4th rater with statistically significant difference ($X^2=59.306$; $p<0.05$). The third text sub-category medians were highest for the 1st and the 7th raters, and lowest for the 4th rater with statistically significant difference ($X^2=48.496$; $p<0.05$). The total text category medians were highest for the 1st and the 7th raters, and lowest for the 4th rater with statistically significant difference ($X^2=55.266$; $p<0.05$). Thus, the findings suggest that the text category scores vary between raters, but do not vary among raters.

The first technical aspects sub-category median was highest for the 3rd rater, and lowest for the 4th and the 5th raters with statistically significant difference ($X^2=157.596$; $p<0.05$). The second technical aspects sub-category medians were highest for the 3rd and the 10th raters, and lowest for the 5th rater with statistically significant difference ($X^2=108.189$; $p<0.05$). The third technical aspects sub-category medians were highest for the 3rd rater, and lowest for the 4th and the 5th raters with statistically significant difference ($X^2=123.992$; $p<0.05$). The total technical aspects category median was highest for the 3rd rater, and lowest for the 5th rater with statistically significant difference ($X^2=145.970$; $p<0.05$). The technical aspects category scores vary between and among raters.

The first content sub-category medians were highest for the 3rd, the 7th and the 8th raters, and lowest for the 4th and the 5th raters with statistically significant difference ($X^2=75.649$; $p<0.05$). The second content sub-category medians were highest for the 1st, the 7th and the 8th raters, and lowest for the 4th rater with statistically significant difference ($X^2=59.156$; $p<0.05$). The total content category medians were highest for the 3rd and the 8th raters, and the lowest for the 4th rater with statistically significant difference ($X^2=63.260$; $p<0.05$). Thus, it can be observed that the content category scores show variety between and among raters.

The first language sub-category median was highest for the 2nd rater, and the lowest for the 4th rater with statistically significant difference ($X^2=58.888$; $p<0.05$). The second language sub-category medians were highest for the 2nd and the 7th raters, and lowest for the 4th and the 5th raters with statistically significant difference ($X^2=74.366$; $p<0.05$). The third language sub-category median was highest for the 1st rater, and lowest for the 4th rater with statistically significant difference ($X^2=40.027$; $p<0.05$). The total language category medians were highest for the 2nd and the 7th raters, and lowest for the 4th rater with statistically significant difference ($X^2=61.078$; $p<0.05$). The language category scores showed variety between and among raters.

The grand total score of the rubric was the highest for 7th rater, and the lowest for 4th rater with significant difference ($X^2=68.249$; $p<0.05$). Thus, the grand total score varied between raters.

Results for the third research question.

The third research question is given below:

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

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. A detailed breakdown of the analysis is presented in the appendix page (see Appendix I). The components of the variance and their relative contribution to the variability of scores are illustrated in Table 13.

Table 13. *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

Table 13 indicates that the largest variance component was attributable to papers (61.8%), which shows that the students displayed different translation performances as measured by the translation tasks. This is a desired outcome since the main aim of an assessment task is to distinguish learners by their writing skills. The second largest variance that contributed to the score variability was the interaction between papers and raters (12.6%), indicating that some raters were inconsistent in their judgments while assigning scores to certain translation papers. The third largest variance was due to the residual (12.1%), which resulted from the interaction of raters, translation papers, quality of translation papers and other unexplained systematic and unsystematic error sources. The fourth greatest variance contributing to the variability of score was due to the interaction between raters and translation quality (12.1%), revealing that the raters differed significantly when assigning scores to different qualities of translation papers. The remaining variance sources such as rater, translation quality and paper by quality were determined to contribute relatively little to the variability in the scores (0%, 0%, and 1.4% respectively).

To identify the sources of variance that contribute to the variability in the scores given to high-quality translation papers, the paper-by-rater (p x r) random effects was carried out. A detailed breakdown of the analysis is presented in the appendix page (see Appendix J). Table 14 displays the components of variance and their relative contribution to the ratings assigned to high-quality translation papers.

Table 14. *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

As Table 14 illustrates, the greatest variance component was attributable to raters (58.5%), showing that the raters were substantially inconsistent while assigning scores to high-quality translation papers. The second largest variance component was determined to be the residual (28.4), indicating that because of interaction between papers, raters and other systematic or unsystematic sources of error, a large variance source cannot be explained in this design. The remaining and smallest variance component following the residual was determined to be papers (13.1%), which means that there was no significant difference in the students' translation skills. Although it is desirable to attain a larger percentage of variance from the papers (object of measurement), the students' contribution to the variability in the scores was relatively low in this design. This might be a result of the homogeneous distribution of students, since high-quality papers were pre-selected specifically for analysis purposes.

The paper-by-rater (p x r) random effects G-study was carried out in order to identify the sources of variance that contribute to the scores given to low-quality translation papers. A detailed breakdown of the analysis is presented in the appendix page (see Appendix K). Table 15 shows the components of variance and their relative contribution to the variability in the scores assigned to low-quality translation papers.

Table 15. *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 15 shows that the greatest variance component was due to raters (59.1%), which indicates that the raters were substantially inconsistent when assigning scores to low-quality translation papers. The second largest variance component after the rater facet was the residual (34.7), revealing that a large source of variance cannot be explained in this design because of the interaction among raters, papers and some other unsystematic or systematic error sources. The remaining and smallest component of variance was due to papers (6.2%), showing that the learners displayed similar translation performances. Again, this might be a result of the homogeneous distribution of pre-selected low-quality papers for the purpose of analysis.

To put this all together, the raters were found to be inconsistent when assigning scores to the translation papers, overall. When the scores given to high- and low-quality translation papers were compared regarding rater leniency and severity, no significant differences were found between these two groups of papers (58.5% and 59.1, respectively). These results indicate that although the students' translation performances within each translation quality group were similar, the raters followed various rating procedures and standards while assessing the papers.

Generalizability and dependability coefficients for ratings.

The calculation of the generalizability and dependability coefficients was performed by applying the random effects G-study design, paper by rater by quality ($p \times r \times q$), to all translation papers, and the random effects G-study design, paper by rater ($p \times r$), to low- and high-quality translation papers separately. Table 16 illustrates the results.

Table 16. *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

The results in Table 16 show that the generalizability and dependability coefficients of all translation papers ($Ep^2 = .96$ and $\phi = .95$) were higher than those of the high- and low-quality translation papers separately ($Ep^2 = .82$, $\phi = .60$ and $Ep^2 = .64$, $\phi = .40$, respectively). The results revealed that although the Ep^2 and ϕ coefficients of the high-quality translation papers were higher than those of the low-quality ones, the coefficients obtained for each quality of paper were far lower than those obtained for all papers.

After the the generalizability and dependability coefficients of the papers had been calculated, various crossed D-study designs were performed on the three sets of translation

papers in order to determine the most appropriate design of scoring in contexts of similar assessment. In a G-study design, increasing the number of facets is assumed to yield higher Ep^2 and ϕ coefficients. For this reason, in the scenarios for high- and low-quality translation papers, the number of raters was increased in order to achieve an acceptable level of Ep^2 and ϕ coefficients, since they were lower in the current scenario. However, raters' number was decreased in the planned scenario for all translation papers since they yielded higher Ep^2 and ϕ coefficients in the current scenario. Table 17 shows the values of the generalizability and dependability coefficients when the number of raters is manipulated in various scenarios.

Table 17. *Generalizability and Dependability Coefficients for All, High and Low Translation Papers*

All translation papers (N= 80, p x r x q)	N of raters	Ep^2	ϕ
	10	.96	.95
	5	.93	.91
	3	.89	.87
	2	.86	.82
High-quality papers (N= 40, p x r)	N of raters	Ep^2	ϕ
	10	.82	.60
	13	.85	.66
	21	.90	.76
	29	.93	.81
Low-quality papers (N= 40, p x r)	N of raters	Ep^2	ϕ
	10	.64	.40
	25	.81	.62
	45	.88	.74
	65	.92	.81

Table 17 illustrates that the Ep^2 and ϕ coefficients were at acceptable levels if the raters' number was decreased to two for all translation papers. As for the high-quality translation papers, the desired Ep^2 and ϕ coefficients could be achieved when the number of raters is increased to twenty nine. For the low-quality translation papers, acceptable generalizability and dependability coefficients would be achieved with a total of 65 raters.

Results for the fourth research question.

The fourth research question is given below:

“Do the variables of gender and rating experience have impact on reliability of the scores assigned to high- and low-quality translation papers?”

Differences in the scores assigned to the translation papers by female and male raters.

In order to understand how raters make decisions while scoring translations of different quality holistically, the effects of the raters' demographic properties on evaluations were examined. Firstly, score differences based on demographic properties for each category of the rubric were analyzed. Table 18 illustrates the results of the Mann-Whitney U test for differences between female and male raters in respect of the scores they gave to the translated texts.

Table 18. *Mann-Whitney U Test Results For Differences Between Female and Male Ratets*

	U	p	Female (Mdn)	Male (Mdn)
Purpose total	64372.500	0.000	0.50	0.70
Culture total	72995.000	0.232	0.50	0.60
Text total	75581.500	0.701	0.30	0.30
Technical aspects total	71145.500	0.076	0.40	0.30
Content total	72866.500	0.217	0.30	0.30
Lang total	71395.500	0.091	0.30	0.30
Grand total	72619.000	0.191	1.20	1.50

According to Table 18, the median of the total purpose category was 0.5 in female raters and 0.7 in males and the difference was statistically significant ($U=64372.500$; $p<0.05$). Purpose scores were higher in male raters than in female raters. The total culture score median was 0.5 in female raters, and 0.6 in males with statistically insignificant differences ($U=72995.000$; $p>0.05$). Culture scores were higher in male raters than in female raters. The text score median was 0.3 in female and 0.3 in male raters with statistically insignificant difference ($U=75581.500$; $p>0.05$). The technical aspects scores were higher in female raters than in male raters. Technical aspects score median was 0.4 in female and 0.3 in male raters with statistically insignificant difference ($U=71145.500$; $p>0.05$). The content score median was 0.3 in female and 0.3 in male raters with statistically insignificant difference ($U=72866.500$; $p>0.05$). The language score median was 0.3 in female and 0.3 in male raters with statistically insignificant difference ($U=71395.500$; $p>0.05$). The total score median was 1.2 in female and 1.5 in male raters with statistically insignificant differences ($U=72619.000$; $p>0.05$). These differences are given in Figure 10.

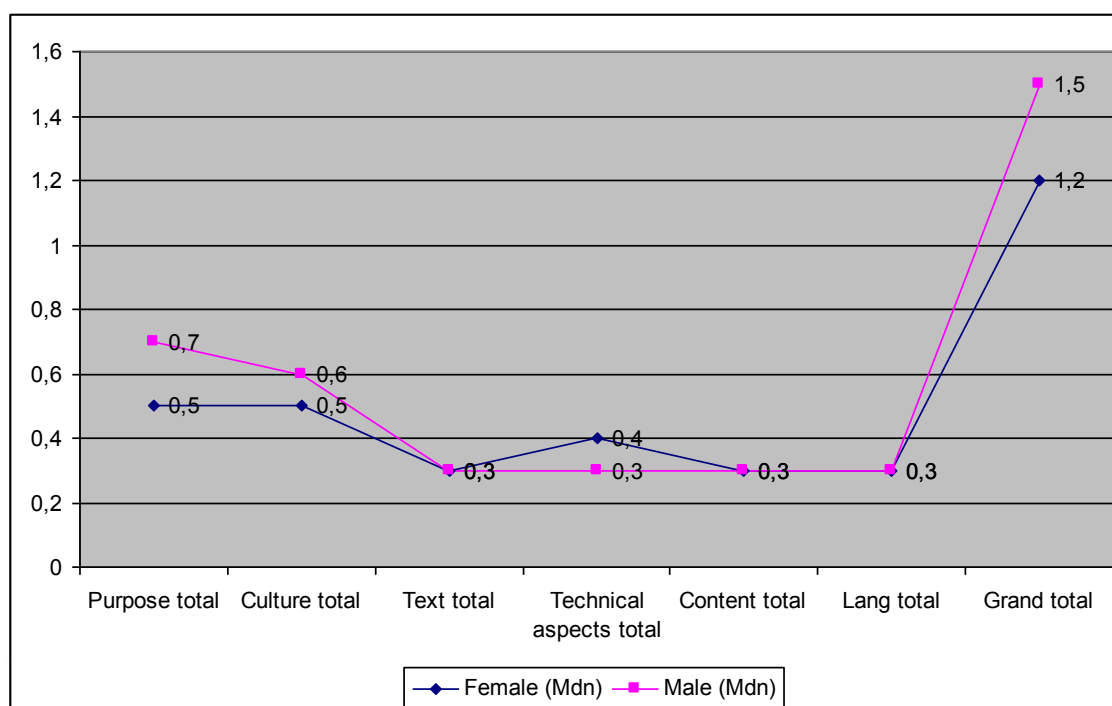


Figure 10. Differences between female and male raters regarding the scores.

The results showed that evaluation differed among the raters only in the purpose category, and gender was not found to be a significant variable in the assessments made by the raters. Table 19 illustrates the generalizability coefficients of the two qualities of translation papers obtained from the G-analysis of the scores assigned to the papers by female and male raters.

Table 19. *G-theory Analysis Results For Differences Between Female and Male Raters*

Translation papers	Number of papers	Female Raters	Male Raters	Ep ² for males	Ep ² for females
High-quality	40	4	6	.67	.72
Low-quality	40	4	6	.47	.48

The results revealed that although higher generalizability coefficients were obtained from female raters for scoring of high- and low-quality papers (.72 and .48, respectively), no significant differences were found between female and male raters in terms of their scoring of high- and low-quality translation papers. A detailed breakdown of the analyses is presented in the appendix page (see Appendix L).

Differences in the scores assigned to the translation papers by experienced and inexperienced raters.

The scores given to each component of the holistic rubric were analyzed to understand whether there were statistically significant differences between the scoring of experienced and

of inexperienced raters. Table 20 shows the results of the Mann-Whitney U test conducted to determine differences between experienced and inexperienced raters.

Table 20. *Mann-Whitney U Test Results for Differences Between Experienced and Inexperienced Rat*

	U	p	Experienced (Mdn)	Inexperienced (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 20, the total purpose median was 0.6 in experienced raters and 0.6 in inexperienced raters with statistically insignificant differences ($U=77563.500$; $p>0.05$). The total culture median was 0.5 in experienced raters, and 0.6 in inexperienced raters with statistically significant difference ($U=71509.000$; $p<0.05$). The total text score median was 0.3 in experienced raters, and 0.3 in inexperienced raters with statistically significant difference ($U=72406.500$; $p<0.05$). Since the distributions were non-normal, it is possible to find significant difference even where the medians are the same. The total technical aspects score median was 0.4 in experienced raters and 0.3 in inexperienced raters with statistically significant differences ($U=71128.500$; $p<0.05$). The total content score median was 0.3 in experienced raters, and 0.4 in inexperienced raters with statistically significant difference ($U=69638.000$; $p<0.05$). The total language score median was 0.3 in experienced raters, and 0.3 in inexperienced raters with statistically no significant difference ($U=69479.500$; $p>0.05$). The total score median was 1.3 in experienced raters, and 1.6 in inexperienced raters with statistically significant difference ($U=70131.000$; $p<0.05$). These differences are illustrated in Figure 11.

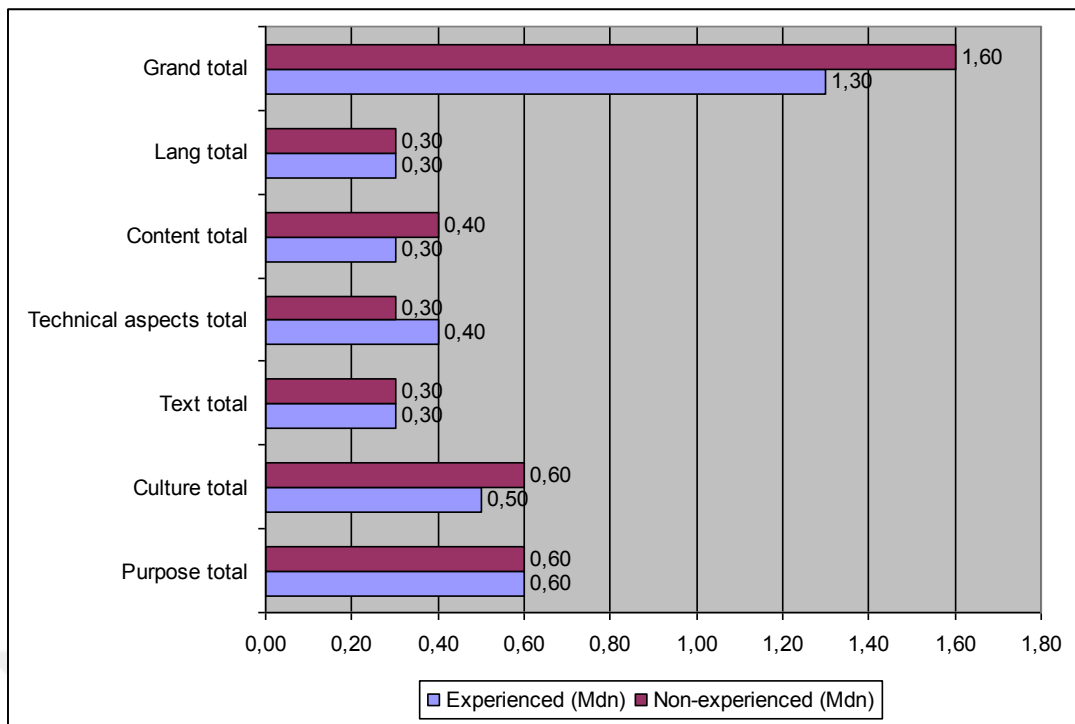


Figure 11. Differences between experienced and non-experienced raters regarding their scorings.

In general, the results showed that inexperienced raters assigned higher scores to the papers than experienced raters did. So, it can be said that experience is a significant parameter affecting raters' scoring of translation papers. Also, calculations of generalizability coefficients were performed for high- and low-quality translation papers to determine differences between experienced and inexperienced raters. The results are given in Table 21.

Table 21. *G-theory Analysis Results For Differences Between Experienced and Inexperienced Ratets*

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

G-study analysis provided higher Ep² and φ coefficients for both high- and low-quality translation papers (Ep²= .80 and .55, and φ=.52 and .32, respectively) rated by inexperienced raters. These results show that inexperienced raters were more consistent in the scores they gave to the papers than experienced raters. A detailed breakdown of the analyses is presented in the appendix page (see Appendix M).

In order to evaluate the score relations between categories of the rubric and to understand raters' trends, correlation analysis was performed. The Spearman's rho correlation analysis results for category and total scores are given in Table 22.

Table 22. *Spearman's Rho Correlation Analysis Results For Task and Total Scores*

		Grand total	Purpose total	Culture total	Text total	Technical aspects total	Content total	Lang total
Grand total	r	1.000	.944**	.979**	.961**	.825**	.973**	.978**
	p		.000	.000	.000	.000	.000	.000
	n	800	800	800	800	800	800	800
Purpose total	r	.944**	1.000	.942**	.892**	.711**	.913**	.928**
	p	.000		.000	.000	.000	.000	.000
	n	800	800	800	800	800	800	800
Culture total	r	.979**	.942**	1.000	.932**	.761**	.955**	.960**
	p	.000	.000		.000	.000	.000	.000
	n	800	800	800	800	800	800	800
Text total	r	.961**	.892**	.932**	1.000	.788**	.940**	.926**
	p	.000	.000	.000		.000	.000	.000
	n	800	800	800	800	800	800	800
Technical aspects total	r	.825**	.711**	.761**	.788**	1.000	.764**	.755**
	p	.000	.000	.000	.000		.000	.000
	n	800	800	800	800	800	800	800
Content total	r	.973**	.913**	.955**	.940**	.764**	1.000	.957**
	p	.000	.000	.000	.000	.000		.000
	n	800	800	800	800	800	800	800
Lang total	r	.978**	.928**	.960**	.926**	.755**	.957**	1.000
	p	.000	.000	.000	.000	.000	.000	
	n	800	800	800	800	800	800	800

The correlation analysis results showed that the highest correlation coefficient was between grand total and culture total ($r=0.979$), followed by language ($r=0.978$), content ($r=0.973$), text ($r=0.961$), purpose ($r=0.944$) and technical aspects ($r=0.825$). The correlations between the categories and the grand total score were highly significant as expected ($p<0.001$). These results show that while culture is the most important category for raters when assessing translation papers, the technical aspects category is the least important category for them.

Results of Qualitative Data Analysis

In this section, the data obtained from the TAPs that were recorded by 8 graders (2 of the raters failed to complete the TAP task during their assessments) were analyzed. A total of 32 TAPs were conducted by each rater in different audio files. Among the translation papers that the raters used for the TAPs, 16 were labeled as high quality translation papers and the other 16 were labeled as low quality translation papers. A total of 256 TAPs were obtained from

the raters for the analysis of qualitative data. Ranging from 4:08 minutes to 14:15 minutes, the TAP recordings were 7:20 minutes long on average. A coding scheme developed by Cumming *et al.* (2002) was applied for the analysis of the data. The scheme was composed of 35 decision-making behaviors that fall into two categories of strategies (interpretation and judgments).

A cumulative approach to translation quality was followed in the qualitative data presentation. For each translation quality, a total of 128 TAPs were recorded by the raters. After the coding task had been completed, a total number of 7.998 strategies were identified. While 4.526 strategies were employed for low-quality translation papers, 3472 strategies were used for high-quality translation papers. This result shows that the raters employed more strategies while assigning scores to low-quality translation papers than while assigning scores to high-quality translation papers. The data obtained from the TAP recordings were used in answering the 5th research question, the findings of which are given below.

Results for the fifth research question.

The fifth research question is given below

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

The last research question deals with how raters make decisions when assigning scores to different quality translation papers holistically. As aforementioned, the raters who took part in this research employed more strategies when scoring low-quality translation papers than when scoring high-quality translation papers. In this section, a more detailed description of the distribution of strategy use by rater will be presented in order to gain an insight into the decision-making behaviors applied for each quality of translation papers.

In order to determine the strategy distribution in respect of type -judgment or interpretation- descriptive statistical analysis was performed. Figure 12 below presents the total number of strategies used by raters and shows how they distribute regarding strategy type and quality of translation papers.

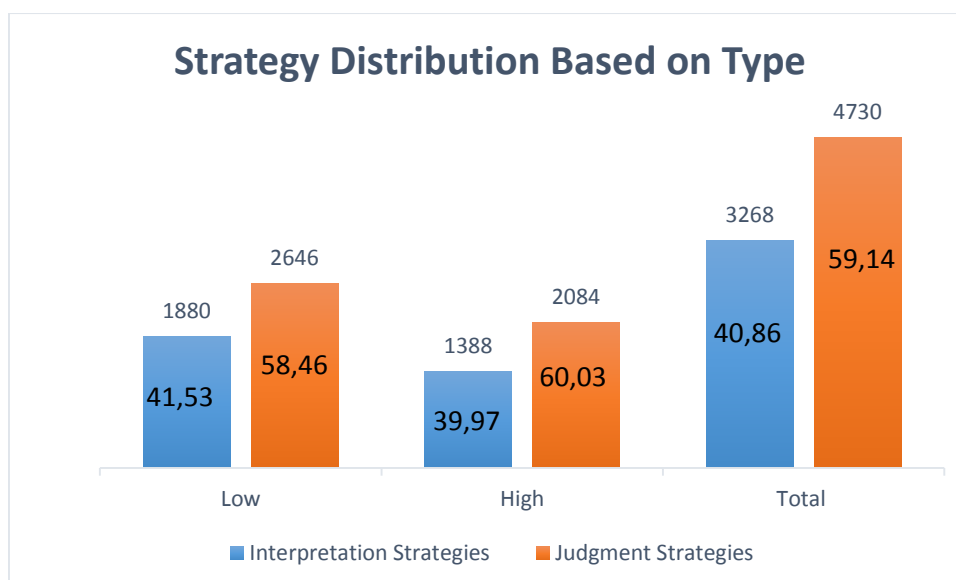


Figure 12. Distribution of strategy type in terms of quality of translation papers.

While 40.86% of the strategies employed by the raters were interpretation strategies, 59.14% of them were judgment strategies. In relation to the high or low quality of translation papers, there is a significant difference (18.28%) between the two types of strategy use. The same trend can be observed for the assessment of both low- and high- quality translation papers. As Figure 9 shows, the raters preferred judgment strategies more while assigning scores to both low-and high-quality translation papers. Then, the decision-making behaviors applied by the raters were divided into three groups based on their focus: 1) self-monitoring focus, 2) rhetorical/ideational focus, and 3) language focus. Figure 13 shows how the strategies distribute in terms of their focus.

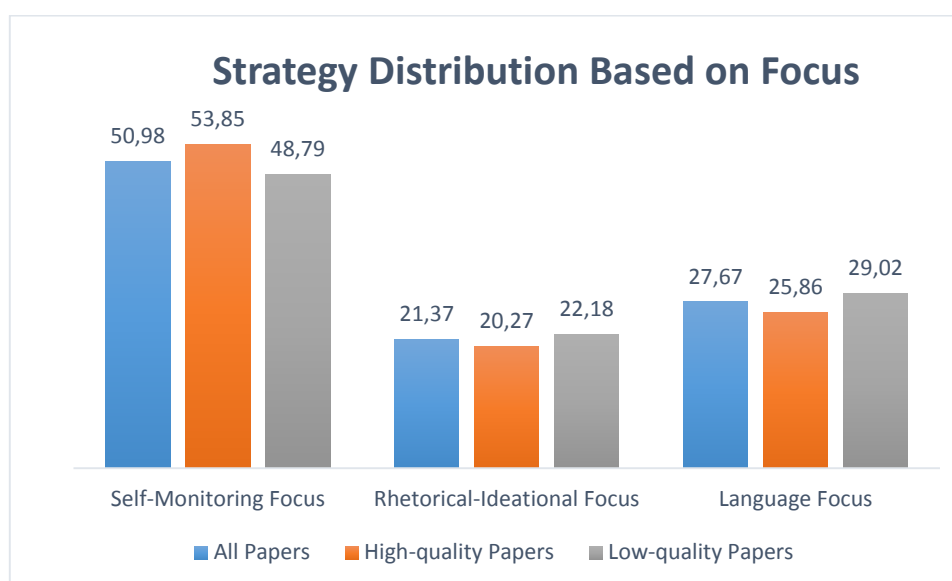


Figure 13. Distribution of strategy focus in terms of quality of translation papers.

According to Figure 13, while assigning scores to all translation papers, raters focused more on self-monitoring decision-making behaviors (50.98%). Secondly, they applied

strategies pertaining to the language focus (27.67%). And lastly, they attended to strategies related to rhetorical/ideational focus (21.37%). The same trend was followed by the raters in evaluations of both high- and low-quality translation papers. However, the raters were observed to apply more language-related strategies for low-quality papers than they did for high-quality papers.

After the examination of strategy use based on focus, the most commonly employed strategies by all raters for all translation papers were identified. Table 23 shows the ranking orders of the decision-making behaviors that the coding scheme included. Then, Table 24 presents a comparison of the frequency of the strategies applied for high- and low-quality translation papers.

Table 23. *The Most Frequently Employed Decision-Making Behaviors by all Raters to all Translation Papers*

Decision-making Behaviors	%
Read or reread text	30.13
Consider spelling or punctuation	6.10
Consider syntax or morphology	5.15
Assess comprehensibility	4.70
Consider own personal response, expectations or biases	4.35
Assess coherence	3.83
Interpret ambiguous or unclear phrases	3.43
Articulate general impression	3.30
Identify redundancies	3.15
Read or interpret scoring scale	2.95
Summarize judgements collectively	2.85
Consider lexis	2.60
Assess task completion or relevance	2.50
Envision personal situation of writer	2.30
Assess reasoning, logic, or topic development	2.20
Assess style, register, or genre	2.15
Edit phrases for interpretation	2.13
Assess text organization	2.10
Rate language overall	1.95
Assess originality	1.78
Articulate or revise scoring	1.75
Scan or skim text	1.65
Consider gravity of errors	1.43
Assess fluency	1.28
Consider error frequency	1.18
Define or revise own criteria	1.15
Classify errors into types	1.00
Compare with other compositions	0.55
Observe layout	0.15
Rate ideas or rhetoric	0.15
Discern rhetorical structure	0.05
Summarize ideas or propositions	0.03
Assess quantity	0.00
Decide on macro-strategy	0.00
Read prompt	0.00
Total	100

As Table 23 indicates, the most frequently applied decision-making behaviors for all translation papers were “read or reread text”, “consider spelling or punctuation”, “consider syntax or morphology”, “assess comprehensibility”, and “consider own personal response, expectations or biases”. These five strategies accounted for 50.43% of the strategies used for all translation papers.

Table 24. *The Most Frequently Employed Decision-Making Behaviors by all Raters to High- and Low-Quality Translation Papers*

High-quality Translation Papers		Low-quality Translation Papers	
Decision-making Behaviors	%	Decision-making Behaviors	%
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

Table 24 compares the distribution of the 35 decision-making behaviors included in the coding scheme for high and low-quality translation papers. As can be seen from Table 24, when the strategies used for two distinct qualities of translation paper are compared, similarities are observed in the rating decision-making behaviors of the raters. Across both qualities of

translation papers, seven of the top ten decision-making behaviors (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 qualities of translation paper, and seemed to focus on language use by relying on strategies of “consider syntax or morphology” and “consider spelling or punctuation”. When considering the remaining eight decision-making behaviors that appeared in the top ten for each quality of translation papers, “consider syntax or morphology” was another commonly preferred strategy, with a rank 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 respectively 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”, ranking 3rd; “read or interpret scoring scale”, ranking 7th ; and “assess task completion or relevance”, ranking 10th. However, 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 for high-quality translation papers. Although these strategies were among the top ten for low-quality translation papers, they had a ranking of 18th, 27th and 17th respectively for high-quality translation papers. Considered as a whole, these decision-making trends indicate that when assigning scores to low-quality translation papers, the raters mostly focused upon the grammar, coherence and comprehensibility of the texts. However, they focused more on identifying redundancies and on the general impression made by the texts while rating high quality translation papers.

Some examples of coded decision-making behaviors applied by raters.

Self-Monitoring Focus-Interpretation Strategies

1. Read or reread text: a) “Hımm, I think I am going to read this sentence again. I might have misunderstood something.” (Elif); b) I can say that this translation paper is bad as a whole, for instance this sentence ‘burnunu kaybetmesine sebep olan nadir kanser şekillerinde’ hımm, the structure of this sentence causes ambiguity.” (Semih)

2. Envision personal situation of the writer: a) “Wow, this is so bad. I don’t understand any of these sentences although they are all in Turkish. I think this student should take Turkish lessons before English ones.” (Ayşe); b) Is this student really majoring in the English Teaching Department? This translation is horrible.” (Hatice)
3. Scan or skim text: a) “When I look at the paper again, I see that there are many spelling errors.” (Tolga); b) “Let me check the punctuation problems I underlined quickly. Well, this paper doesn’t deserve a high score for this category.” (Meriç)
4. Read or interpret scoring scale: “The first category on the rubric deserves a high score since this student translated in accordance with the terminology” (Ülfet); b) “The target text is structured cohesively (reading a sub-category loudly), no, I don’t think so. That’s why, I’m giving 0.1 point to this sub-category.” (Selahattin)

Self-Monitoring Focus-Judgment Strategies

1. Consider own personal response, expectations, biases: “I think that fluidity is necessarily the main goal for a good translation. Also, I think a good translation is never enough. Maybe I’m too harsh, I don’t know” (Ayşe); b) “I expect a good translation to be bold and creative, but I can’t see any of these features in these papers.” (Tolga)
2. Define or revise own criteria: a) “Hmm, these last sentences changed my mind about the coherence of the translation.” (Elif); b) At first, I thought this was a good translation. Now, I’ve changed my mind.” (Tolga)
3. Compare with other compositions: a) “The language used in this translation is more fluent compared to the others, as if it was translated by a professional.” (Hatice); b) “Well, this is the worst translation I have evaluated.” (Semih)
4. Summarize judgments collectively: a) “This is generally a good translation. There are a few spelling mistakes and just a few punctuation problems. Considering it as a whole, I can easily get the message of the original text” (Tolga); b) “These expressions are all problematic. There are serious grammatical problems in nearly all of these sentences. Too many spelling and punctuation mistakes, no coherence, and no transition between the sentences” (Ülfet)
5. Articulate general impression: “This translation is good, in fact.” (Semih); b) This translation can be classified as a bad one.” (Meriç)
6. Articulate or revise scoring: “I am giving 1 point to the purpose category” (Selahattin); b) “Overall, this translation got 8 out of 10 points.” (Ayşe)

Rhetorical and Ideational Focus-Interpretation Strategies

1. Interpret ambiguous or unclear phrases: a) “What are you trying to say in this sentence? I really don’t understand. This is just ridiculous.” (Elif); b) “Hmm, I think the student could not fully understand what this sentence in the original text wanted to say. That’s why, there is confusion in this sentence” (Meriç)

Rhetorical and Ideational Focus-Judgment Strategies

1. Assess reasoning, logic or topic development: a) “Although there is no incoherency in these sentences, some details are missing” (Tolga); b) “Hmm, the things written in this sentence are not logical. Well, this sentence doesn’t make sense at all.” (Selahattin)
2. Assess task completion or relevance: a) “This translation has nothing to do with the original text. The message that the original text wanted to convey was never understood by this student.” (Ülfet); b) “No, this sentence is not related to the original one.” (Hatice)
3. Assess coherence: a) “Let me check. Well, I don’t see anything like that in the original text. This expression in the translation contradicts the one in the source text.” (Elif); b) “There is no coherence in this translation.” (Tolga)
4. Assess originality: a) “This was an original translation, I liked the language used here.” (Meriç); b) “Translation of this sentence is really good and not like the others did.” (Tolga)
5. Identify redundancies: a) “This phrase is unnecessary, I can’t see anything like that in the original text”. (Ayşe); b) “It feels like this student made up some things that are not mentioned in the original text”. (Selahattin)
6. Assess text organization: “This translation paper has no organization. I do not understand which conversation belongs to whom.” (Semih); b) “There is no organization in this paper.” (Elif)
7. Assess style, register or genre: a) “The semiotic functions of the original text were not taken into consideration here.” (Ayşe); b) “The linguistic choices made in this translation are problematic.” (Tolga)
8. Rate ideas or rhetoric: a) “I liked the idea of the student using this phrase here.” (Meriç); b) “This student found interesting Turkish equivalents to proper English nouns in the original text.” (Tolga)

Language Focus-Judgment Strategies

1. Observe layout: a) “When I look at the translation paper, I see that the student didn’t use paragraph indents” (Elif); b) “This paper has no organization, quotation marks are missing and paragraphs are mingled here.” (Semih)
2. Classify errors into types: a) “This student made many spelling and punctuation mistakes.” (Selahattin); b) “I see subject verb agreement problems in these sentences.” (Hatice)
3. Edit phrases for interpretation: a) “Haa, this student must have forgotten to use ‘sonra’ in this expression.” (Elif); “A word is missing in this sentence. The student might have forgotten to use ‘için’ here.” (Semih)

Language Focus-Judgment Strategies

1. Assess comprehensibility: a) “I don’t understand what you are talking about. This doesn’t make sense.” (Meriç); b) “I find it difficult to understand what is written here.” (Tolga)
2. Consider frequency of errors: “There are lots of spelling and punctuation errors in this paper.” (Selahattin); b) “This is the fourth sentence that has no subject verb agreement.” (Semih)
3. Consider gravity of errors: a) “This paper is full of spelling mistakes.” (Ayşe); b) “Because of the punctuation mistakes, I can’t understand any of the expressions here.” (Semih)
4. Assess fluency: a) “There is no flow between the sentences.” (Elif); b) “Transitions between the expressions, I mean between the sentences are problematic.” (Tolga)
5. Consider lexis: a) “The student used ‘utanmaz aktör’, hmm it is not an adjective describing the actor. It is the name of the series he is acting in.” (Elif); b) “It says ‘traktörü doldurdu’ here. The correct translation of the word should be ‘kamyonet’.” (Ayşe)
6. Consider syntax or morphology: a) “I’ve noticed some syntax problems in this paper.” (Ülfet); b) “I don’t understand what it says here. I guess it is because of syntax problems.” (Hatice)
7. Consider spelling or punctuation: a) “There should have been a quotation mark here.” (Semih); b) “This paper has many spelling errors.” (Ayşe)
8. Rate language overall: a) “The language of the translation is generally fluent, I had no difficulty in reading it.” (Semih); b) “Overall, choice of phrases is good in this paper, and it has a fluent language.” (Elif)

Summary of the Findings

In this section, a summary of the findings obtained from the analyses of SPSS, Edug and TAPs for each research question are presented. Table 25 presents a summary of the results belonging to each research question.

Table 25. *Summary of the Findings Pertaining to Each Research Question*

Research Questions	Results
RQ1. Are there any significant differences among the holistic scores assigned to low- and high-quality translations?	- Yes. The holistic scores given to two distinct qualities of translation paper differed significantly. Raters gave higher scores to high-quality translation papers. - the largest variance component was quality of the papers (%36.4). The second greatest variance was because of the interaction between rater and quality (%35.1), suggesting that raters showed significant differences in the scores they gave to two different qualities of paper. The third largest variance resulted from raters (%20.4) indicating that some of the raters were inconsistent regarding their judgements while assigning scores to some certain translation papers.
RQ2. What is the score variability between and among raters' scores assigned to each category and sub-category of the holistic rubric?	- Regarding six components of the scale, the results obtained from the Kruskal Wallis Test (68.249) and G-theory variance analysis (%59.4) showed statistically significant differences between and among raters. - In general it was found that there was not a linear or regular evaluation trend among raters. - Regarding grand total scores, 7th rater evaluation was the highest and 4th rater evaluation was the lowest among other rater scores with significant differences. This finding suggests that while the 4th rater adapted a strict trend in assigning scores to the papers, the 7th rater followed a generous trend in assigning scores to the papers. - Results obtained from the Kruskal Wallis test indicated that scores assigned to the sub-categories of the purpose, culture and text components varied significantly between the raters, but not among them. On the other hand, scores given to the sub-categories of the technical aspects, content and text components revealed significant variation both between and among raters.
RQ3. What are the sources of score variation that contribute to the score variability of the holistic scores assigned to high- and low-quality translation papers?	- G-study variance analysis performed to all translation papers to identify score variability sources indicated that papers (61.8%) was the main variance component contributing to the variability. The other variance components contributing to the score variability were found as follows: interaction between papers and raters (12.6%), residual (12.1%) and interaction between raters and paper quality (12.1%). - G-study variance analysis results obtained from high-quality translation papers revealed that the largest variance component was attributable to raters (58.5%). Other components that contribute to score variability are residual (28.4%) and papers (13.1%) respectively. - Lastly, G-study variance analysis results obtained from low-quality translation papers showed that the greatest variance component was attributable to raters (59.1%). Other components contributing to score variability were determined as residual (34.7%) and papers (6.2%) respectively.

Table 25. (continued)

RQ4. Do the variables of gender and rating experience have impact on reliability of the scores assigned to high- and low-quality translation papers?	- Statistically no significant difference was determined between male and female raters regarding the scores they assigned to each category of the rubric. Thus, gender was not found to be a significant variable impacting assessments of the raters. - Experience was determined as a significant parameter affecting raters' scoring of translation papers since G-study analyses yielded higher Ep^2 coefficients for both qualities of translation paper rated by inexperienced raters. Thus, results showed that inexperienced raters were more consistent in the scores they assigned to the papers than experienced raters did.
RQ5. How do raters make decisions while assigning scores to different quality of translation papers holistically?	- Raters were found to prefer judgement strategies more while assigning scores to both qualities of translation paper. Considering focus of the strategies, they applied more self-monitoring decision-making behaviors (50.98%) compared to the language (27.67%) and rhetorical/ideational strategies (21.37%). - The most frequently applied decision-making behaviors for all translation papers were determined as "read or reread text", "consider spelling or punctuation", "consider syntax or morphology", "assess comprehensibility", and "consider own personal response, expectations or biases". - For high-quality translation papers, the most frequently used decision-making behaviors were found as "read or reread text", "consider spelling or punctuation", "articulate general impression", "consider syntax or morphology", and "consider own personal response, expectations or biases". - For low-quality translation papers, the most frequently applied decision-making behaviors were determined as "read or reread text", "consider spelling or punctuation", "assess comprehensibility", "consider syntax or morphology", and "interpret ambiguous or unclear phrases". - Decision-making trends of the raters suggested that when assigning scores to low-quality translation papers, they mostly centered upon grammar, coherence and comprehensibility of the texts. However, they focused more on identifying redundancies and on general impression of the texts while scoring high-quality translation papers.

CHAPTER V

Discussion and Conclusion

This section begins with a discussion of the results yielded from the quantitative and qualitative data analysis with an emphasis on the impact of paper quality on the reliability of holistic scores. Then, the chapter continues with a discussion of the implications from a methodological and pedagogical perspective. Lastly, recommendations for future research are presented.

Discussion

This thesis study examined the scores assigned to translation papers by raters and the impact of translation quality on the holistic scores of translation papers through a mixed-methods analysis. Furthermore, this research investigated the decision-making behaviors of the raters while assigning scores to two distinct qualities of translation paper. This chapter discusses the findings pertaining to the research questions in detail. Following the discussions of the results obtained from the analysis of the quantitative data set in response to RQ1, RQ2, RQ3 and RQ4, the findings of the qualitative data analysis for RQ5 are discussed.

Discussion of findings for RQ1.

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

The translation papers used in this research were obtained from second year university students majoring in English in an ELL Department at a state university. Then, for the purpose of the study, the collected papers were grouped into two distinct qualities (high and low). The raters who participated in the research, were working as instructors in ELL and EFL Departments at state universities. Before the rating procedure, the raters were provided with necessary information regarding background and the proficiency level of the students.

Although the instructors were not informed in advance regarding the division of quality in the translation paper pack, they appeared to be able to assign different grades to two different qualities of translation paper. 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 higher or lower grades to a paper

(Freedman, 1981). Furthermore, since the raters were aware that the texts were translated by EFL students, they might have had high expectations regarding their performance. In this regard, the raters might have assigned lower scores to low-quality translation papers and to some of the high-quality translation papers. In a study conducted by Baker (2010), raters were found to make a distinction between high- and low-quality papers. However, they showed a tendency to give different grades to the same papers under different conditions.

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

The participating raters were able to detect translation papers with richer textual features and assigned higher grades to these papers. This finding coincides with the results of the previously conducted studies (Ferris, 1994; Han, 2017; Huang *et al.*, 2014; Vaughan, 1991). However, raters were not able to agree on the level of translation papers with poor textual features, since more variation was identified in the scores they gave to low-quality papers. This finding contradicts those of Şahan (2018) in his doctoral dissertation study. In his study, the researcher determined that raters were able to agree on papers with poor text features better, since there was less variation in the scores they assigned to low-quality papers in comparison with high-quality papers. Again, in a study conducted by Mendelsohn and Cumming (1987), the researchers found that raters agreed on rating while evaluating low and high-quality compositions, but that the scores they assigned to middle-quality compositions showed significantly higher variation. In this study, the variation among the raters' scores shows that there is a fairness problem in assessment regarding the writings of low-proficiency students. This might be associated with the beliefs of the raters regarding a low-quality translation. Also, the emotions, biases and expectations of the raters might have affected the scores they gave to low-quality papers.

In brief, quality of translation paper was found to have an impact on the raters' scoring behaviors. Firstly, the participating raters were able to distinguish between low- and high-quality translation papers. Secondly, although the scores that the raters gave to high-quality translation papers were similar; they differed in the scores they gave low-quality translation

papers. Lastly, the raters were found to be more consistent when rating high-quality papers, suggesting that the inter-rater reliability of the high-quality papers was higher than that of the low-quality ones.

Discussion of findings for RQ2.

What is the score variability between and among raters' scores assigned to each component and sub-component of the rubric?

Regarding the six components of the holistic rubric used in the study, the scores assigned to the purpose category varied significantly between and among the raters, ranging from 0.40 to 0.80. While the 2nd rater's rating was the highest for this category, the 5th rater's rating was the lowest. Similarly, the scores assigned to the culture category showed variability between and among the raters. While the 7th rater's rating was the highest for this category (1.40), the 5th rater's rating was the lowest (0.65), indicating significant difference. For the text category, the 1st and the 7th raters' ratings were the highest and the 4th rater's rating was the lowest, showing significant difference. Again, the scores assigned to the technical aspects category varied between and among the raters, with the highest mean (1.40) for the 3rd rater and the lowest mean (0.75) for the 5th rater. Regarding the content and language categories, score variability was significant between and among the raters, with the highest means for the 2nd and 7th raters and the lowest means for the 4th rater.

In brief, these results indicated that the 7th rater's rating was the highest and the 4th rater's rating was the lowest among the raters' scores and that there were significant differences between them. The results revealed that there was not a linear or regular rating trend among raters, and scores varied between and among raters for each category of the rubric. This finding suggests that the dissimilar ratings among raters might be related to how they understood and interpreted the given scale and to the criteria to which they gave priority. Thus, it can be said that if raters are not well-trained in how to make use of a given scoring scale, the use of a rubric might not differ from assessments carried out with no specific criteria (Rezaei & Lovorn, 2010). In this study, although the raters were provided training in how to use the given holistic rubric, diversity in scores assigned to translation papers could not be eliminated to a large extent, and this finding parallels what Esfandiari and Myford (2013) found in their study. Also, inconsistency among raters' scores assigned to similar student writings might result from differences in rater severity and perspective (Lunz, Wright & Linacre, 1990). In much previously conducted research (De Gruijter, 1984; Littlefield, Harrington, Anthracite & Garmon, 1981; Lunz *et al.*, 1990), it has been argued that the application of a standardized

scoring scale is useful in defining the evaluation process, but that it might not work for removing differences in the perspective and severity of the raters.

Discussion of findings for RQ3.

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

The results obtained from generalizability analysis in a design of $p \times r \times q$ performed on all translation papers revealed that the largest variance (61.8%) was attributable to papers (p). This indicates that students, as desired, 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 high-quality papers (13.1%) in comparison to the variance observed in low-quality papers (6.2%). Since there were more homogeneous student groups in each of these 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 the students in the higher proficiency group displayed more dissimilar translation abilities from each other than did those compared to the ones in the lower proficiency group. This finding might also 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 a design of $p \times r \times q$ 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 due to the residual (12.1%), which was a result of the interaction among raters, papers, translation quality and other systematic or unsystematic error sources. 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 displayed similar trends in respect of leniency and severity and when assigning grades to high- and low-quality papers (58.5% and 59.1%, respectively).

However, in the $p \times r \times q$ 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 a large-scale assessment context where students' proficiency levels vary (Şahan, 2018). In other words, when raters are given a pile of low-quality papers, they tend to be more inconsistent in their scoring. Therefore, instructors might take advantage of rater-training programs to conceptualize constructs of quality in students' writings. 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 the raters varied considerably in the scores they gave to two distinct qualities of paper. The other sources of variance such as quality (0.0%) and the interaction between paper and quality (1.4%) did not seem to have a significant impact on the score variability. These findings parallel those of Şahan (2018) in his thesis study.

In this study, large variances were determined in the rater facet and the residual. Although the ratings of mixed-quality yielded papers high dependability coefficients ($\phi=.95$), the ratings of high-quality and low-quality translation papers yielded low dependability coefficients ($\phi=.60$ and $\phi=.40$, respectively). This result might be related to number of raters because, when the number of facets is increased, higher dependability coefficients are likely to be obtained (Brennan, 2001; Güler et. al., 2012; Şahan, 2018). Thus, in order to achieve acceptable values for the dependability coefficients, various crossed design D-studies were performed on the three sets of translation papers. For the mixed-quality translation papers, the raters number was decreased in the planned scenario since it yielded high E_p^2 and ϕ coefficients in the current scenario. For this reason, decreasing the number of raters to two yielded acceptable coefficients of dependability. However, for the scorings of high-quality translation papers, the desired dependability coefficients could be achieved when the number of the raters was increased to twenty-nine. For low-quality translation papers, acceptable generalizability and dependability coefficients would be achieved with a total of 65 raters. When two qualities of translation paper are scored separately, reducing the number of the raters would yield lower dependability coefficients, which might raise concerns regarding the fairness of the scores (Huang, 2008). These results indicate that the inclusion of two or more raters in the process of assessment of students' written performance is necessary for fairer scoring, especially in tertiary education. It is recommended that the number of tasks should be increased in order to achieve higher reliability in scorings since the assessment of writing samples by students belonging to the same level of proficiency might be problematic (Baker, 2012). Therefore, when

assessing the writing ability of similar student groups, using various tasks and topics might be an efficient way of maximizing the fairness of judgments.

To summarize, when students with low and high proficiency levels were considered collectively, the translation task separated them in regards to their abilities in translation. When two qualities of translation paper were examined individually, the fluctuation among the scores given to low-quality papers was substantially greater than the fluctuation among the scores of high-quality papers. In all three designs, the contributions of the residual facet were determined to be large, indicating the impact of some unknown facets on the variability of scores. Lastly, the variance attributable to the interaction between raters and paper quality suggested that some of the raters varied considerably in their ratings while assigning scores to different qualities of papers.

Discussion of findings for RQ4.

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

A Mann-Whitney U test was conducted to determine the differences between male and female raters regarding the scores they assigned to each category of the holistic rubric. Among the six categories of the holistic rubric (purpose, culture, text, technical aspects, content and language), male and female raters differed substantially in the scores they assigned to the first category (purpose). While the score median of the purpose category was 0.5 in female raters, it was 0.7 in male raters ($U=64372.500$; $p<0.05$). For this category, male raters were observed to assign higher scores than female raters did. Regarding the remaining five categories of the holistic rubric, no significant differences were determined between female and male raters. While the total median score of all the categories was 1.2 in female raters, it was 1.5 in male raters ($U=72619.000$; $p>0.05$). Also, G-analysis was conducted to determine the generalizability coefficients of high- and low-quality translation papers scored by female and male raters. The results showed that although higher generalizability coefficients were obtained from female raters in the scoring of high- and low-quality papers (.72 and .48, respectively), no significant differences were found between female and male raters in respect of their scoring of high- and low-quality translation papers. Thus, it can be said that, although the difference was not significant between male and female raters, male raters tended to assign higher scores to the translation papers in comparison to the female raters. These results coincide with the findings of a study carried out by Soler (2000). Investigating the scoring behaviors of raters, the researcher found that female and male raters displayed similar rating behaviors while assigning scores to student essays. Also, in a small study on rater impact on score variability,

Carroll (1991) found that although the difference was not significant, male raters assigned higher scores to the candidates than female raters did.

A Mann-Whitney U test was carried out to examine the scores given to each category of the holistic scale and to determine whether there were statistically significant differences between the scores of experienced and inexperienced raters. Analyses of the scores indicated that experienced and inexperienced raters differed substantially in the scores they gave to the culture, technical aspects and content components. When the scores given to all components of the rubric were considered collectively, the total score median was 1.3 for experienced raters and 1.6 for inexperienced raters, and the difference between them was statistically significant ($U=70131.000$; $p<0.05$). When raters' prior rating experience was taken into consideration, it was determined that the experienced raters gave lower scores to the translation papers than the inexperienced raters did. This finding indicates that while the experienced raters scored the translation papers more leniently, the inexperienced raters assessed them more severely. Similarly, previously conducted research found that, although raters with different rating experience produced more similar analytic scorings, experienced raters tended to give higher holistic scorings (Song & Caruso, 1996). Likewise, the findings of the current study support those of other research that has found that more experienced raters are more lenient than inexperienced raters (Shirazi, 2019; Şahan, 2018; Weigle, 1999). On the other hand, findings of this study contradict the results of some previous research which found that inexperienced raters tended to give higher scores (Rinnert & Kobayashi, 2001) and to assess papers more leniently (Barkaoui, 2011; Sweedler-Brown, 1985). However, in a study carried out by Shirazi (2019), both experienced and novice raters were found to perform quite alike in terms of leniency and severity. These results suggest that the ways in which raters interpreted the given rubric and the criteria to which they gave priority might have influenced the dissimilarity in the ratings. Thus, it can be said that if raters are not trained in how to use the given scale efficiently, the use of a rubric might not yield scorings different from those of criteria-free evaluations (Rezaei & Lovorn, 2010).

Also, calculations of generalizability coefficients were performed on high- and low-quality translation papers in order to see whether there were any differences between experienced and inexperienced raters. In this context, the high-quality papers scored by inexperienced raters yielded higher reliability coefficients than those scored by more experienced raters (.80 and .46, respectively). Similarly, regarding the scorings of the low-quality papers, the reliability coefficients obtained from inexperienced raters were found to be higher than those of more experienced raters (.55 and .13, respectively). These results suggest

that in the ratings of both high- and low-quality translation papers, inexperienced raters were more consistent than more experienced raters. Thus, the inconsistency that more experienced raters displayed was found to be substantially higher than that of inexperienced 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 dependent on their expectations (Eckes, 2008). This may also explain why experienced raters tended to give relatively higher scores to the papers overall. It is possible that more experienced raters adhere to the evaluation criteria they have already adopted, since it is difficult for them to get used to the new criteria offered to them (Cumming *et al.*, 2002). When the expectations of raters change according to their prior rating experience, the scores they assign to papers differ as well. For this reason, it is likely that higher inconsistency will be observed among more experienced raters. That is to say, more experienced raters may show a greater tendency to rely on their self-criteria than inexperienced raters do (Şahan, 2018). However, in a study conducted by Cumming (1990), high-experienced raters were found to be more consistent in the scores they assigned to student essays, in contrast to the findings of this study. Unlike Cumming's research (1990), this study did not focus on score differences between novice and expert assessors; instead, raters with a variety of prior rating experience were examined here.

To summarize, contrary to the variable of gender, rater experience was determined to play a significant role in the variability of scores given to translation papers. Firstly, although male and female raters differed substantially in the scores they assigned to the first category (purpose), no significant differences were determined between them regarding remaining five categories of the holistic rubric. Also, the G-analysis results showed that, although higher generalizability coefficients were obtained from female raters for the scoring of high- and low-quality paper, no significant differences were found between female and male raters in terms of their ratings of the two distinct qualities of translation paper. Thus, gender was not found to be a significant variable impacting upon the reliability of the scores assigned to high- and low-quality translation papers. Secondly, regarding previous experience of rating, it was determined that experienced raters assigned considerably lower scores to the translation papers than the inexperienced raters. Analyses of the scores given to the categories of the holistic rubric revealed that experienced and inexperienced raters differed substantially in the scores they gave to the culture, technical aspects and content components. Also, high-quality papers scored by inexperienced raters yielded higher reliability coefficients than those scored by more experienced raters. Similarly, regarding the scorings of the low-quality papers, the reliability coefficients obtained from inexperienced raters were found to be higher than those of more experienced raters. Furthermore, more experienced raters were more inconsistent in the scores

they gave to both qualities of paper. Additionally, inexperienced raters were found to be more consistent while assigning scores to high-quality papers than to low-quality papers.

Discussion of findings for RQ5.

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

The data collected from the TAPs were used in coding the strategies that the participating raters applied during the assessment of translation papers. Collectively, the results revealed that the raters applied significantly more judgment strategies (59.14%) than interpretation strategies (40.86%). While this finding coincides with what Barkaoui (2010) found, it contradicts the findings of Cumming *et al.* (2012), Gebril and Plakans (2014) and Şahan (2018). It was found that the same trend in strategy use was followed by the raters for both low- and high- quality translation papers. Also, the raters who participated in this study employed more strategies when grading low-quality translation papers in comparison to high-quality translation papers, indicating that the assessment of low quality translation papers was more challenging for the raters in comparison to assessment of high quality papers. In addition to that, the raters were observed to apply slightly more interpretation strategies to low-quality translation papers than to high-quality translation papers; and they applied slightly more judgment strategies to the latter, which is also consistent with Şahan (2018). These slight differences show that quality of paper has an impact on the decision-making behaviors that raters prefer to adopt.

Considering strategy use based on focus, the raters mostly preferred strategies with a self-monitoring focus (50.98%), followed by those with a language focus (27.67%) and a rhetorical/ideational focus (21.37%), respectively. The raters were observed to follow the same trend for both high- and low-quality translation papers. In previously conducted studies (Barkaoui, 2010; Gebril & Plakans, 2014; Şahan, 2018), it was found that the strategies most frequently applied by raters were related to self-monitoring focus, while rhetorical/ideational strategies were preferred secondly and language-oriented strategies were last. Also, in a study carried out by Han (2017), raters were found to prefer mostly language focus strategies, followed by rhetorical and ideational strategies, and self-monitoring strategies, respectively. In addition, the raters in this study were found to use more language-related strategies while assessing low-quality translation papers compared to high-quality papers. Similarly, in studies conducted by Cumming *et al.* (2002), Gebril and Plakans (2014) and Şahan (2018), raters were found to apply more language-related strategies to low-quality translation papers than to high-

quality ones. This result might be due to the fact that the low quality translation papers included more language-related problems than did high quality papers.

The most frequently applied strategy for all translation papers was “read or reread text” (30.13%), followed by “consider spelling or punctuation” (6.10%) and “consider syntax or morphology” (5.15%). The “read or reread text” strategy was expected to be applied 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 previously conducted research (Barkaoui, 2010; Şahan, 2018). The fact that the writing task was translation might have caused the raters to concentrate on the appropriate use of syntax and morphology in the translation papers and on whether the spelling and punctuation were in accordance with the target language. These three decision-making behaviors constituted 41.38% of all the strategies employed by the raters. The fourth most frequently applied decision-making behavior was “assess comprehensibility” (4.70%), showing that raters mostly struggled to understand what what was meant to be expressed in the translation papers. Given that the raters first read the original texts and interpreted them in their minds, it is not surprising that they expected to see whatever they had in their minds on the translation papers. “Consider own personal response, expectations or biases” (4.35%) was the fifth most commonly applied decision-making behavior for all papers, indicating that raters mostly relied on their intuitions and beliefs while making judgments about the papers. In addition, frequent use of this strategy has been said to account for variability in holistic scores and for leniency and severity differences among raters (Mcnamara & Roever, 2006). The sixth most preferred behavior by the raters was “assess coherence” (3.83%), indicating that the raters focused on the flow of ideas and on whether the expressions in the translations were logical or not. The seventh most commonly used strategy for all papers was “interpret ambiguous or unclear phrases” (3.43%), which indicates that raters put more weight on language-related problems. “Articulate general impression” (3.39%) ranked eighth in order, indicating that the raters mostly tried to get a general impression from the papers they read. As Lumley claims, “the raters’ overall impression, based on their professional judgment, is the primary influence of their assessments, rather than attention to the scoring categories as discrete entities, and that raters employ various strategies to ensure that their scores reflect this impression” (2000, p.281). “Identify redundancies” (3.15%) was the ninth most common decision-making behavior, which shows that raters determined the use of unnecessary phrases in the translations that were not included in the original texts. Lastly, “read or interpret scoring scale” was the tenth most commonly applied strategy by the raters. This result is not surprising, since all the raters were expected to apply the rubric for scoring after reading each translation paper (Şahan, 2018).

When the decision-making behaviors used for low- and high-quality translation papers were compared, similarities were observed in the rating strategies of the raters. Across the two qualities of translation papers, seven of the top ten decision-making behaviors (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 qualities of translation paper, and seemed to focus on language use by relying on strategies of “consider syntax or morphology” and “consider spelling or punctuation”. These findings coincide with what Gebril and Plakans (2014) found in their study. The strategy of “consider syntax or morphology” ranked 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. This shows that raters had more difficulty in understanding low-quality texts than high-quality texts. Likewise, while the strategy of “assess coherence” ranked 6th for low-quality translation papers, it ranked 9th for high-quality translation papers, showing that the raters centered more upon shifts in coherence while scoring low-quality translation texts. While the strategies of “consider own personal response, expectations or biases” and “identify redundancies” ranked 5th and 6th respectively for high-quality translation papers, they ranked 7th and 10th respectively for low-quality translation papers. These findings indicate that the raters displayed different degrees of severity and leniency while scoring high- and low-quality translation papers, even though they had been given training in following the criteria in the holistic rubric. Similarly, in a study on rater bias, Schaefer (2008) performed a rater and criterion interaction analysis and found significant variations in the scores that raters gave to sub-categories of the scale used in the study.

Other strategies that appeared among the top ten most frequently used for high-quality translation papers were “articulate general impression”, ranking 3rd; “read or interpret scoring scale”, ranking 7th; and “assess task completion or relevance”, ranking 10th. However, these strategies had rankings of 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 commonly employed for low-quality translation papers than for high-quality translation papers. Although these strategies had rankings among the top ten for low-quality translation papers, they had rankings of 18th, 27th and 17th respectively for high-quality translation papers. These results show that while scoring high-quality translation papers, the raters concentrated on their general impression and evaluated the relevance of the message given in translations to the message in the original texts. On the other hand, while rating low-quality papers, the raters mostly centered upon ambiguity in meaning and characteristic features of the translator. Considered collectively, these decision-

making trends suggest that when assigning scores to low-quality translation papers, the raters mostly centered upon the grammar, coherence and the comprehensibility of the texts. However, they focused more on identifying redundancies and on the general impression given by the texts while rating high quality translation papers.

To summarize, the raters in this study were found to apply significantly more judgment strategies than interpretation strategies while assigning scores to all translation papers. Also, they employed significantly more strategies while grading low-quality translation papers when compared to high-quality translation papers. Regarding the use of strategies based on focus, the raters were found to prefer mostly strategies with a self-monitoring focus, followed by a language focus and a rhetorical/ideational focus, respectively. They followed the same trend for both high- and low-quality translation papers. In addition, they used more language related strategies while assessing low-quality translation papers. When the decision-making behaviors used for low- and high-quality translation papers were compared, similarities were observed in the rating strategies of the raters. Across the two qualities of translation papers, seven of the top ten decision-making behaviors (with a different ranking order) employed by the raters were found to be the same. Further, the raters displayed different degrees of severity and leniency while scoring high- and low-quality translation papers, although they were provided with training in following the criteria in the holistic rubric. Overall, the discussion in this chapter presents findings that account for fluctuation between the scorings of raters, which might be useful in terms of gaining a deeper insight into variability in scoring, and in the performance of a more extensive rater training procedure.

Conclusion

First, in this research, two different qualities of translation paper (high and low) were used and no information was provided to the raters regarding this distinction. The statistical analyses employed showed that the raters differed substantially in the scores they gave to low-quality and high-quality translation papers, indicating that they were able make a distinction 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. Thus, the inter-rater reliability of the high-quality translation papers was determined to be higher than that of the low-quality ones.

Second, the Kruskal Wallis test performed to determine score variability between and among the scores that the raters gave to each coategory of the scale showed substantially significant differences. The results indicated that the 7th rater's rating was the highest and the

4th rater's rating was the lowest among the raters' scores and that the difference was statistically significant. In general, it was found that there was not a linear or regular assessment trend among raters, and that scores varied between and among raters for each category and sub-category of the rubric. Therefore, even though the raters were trained in how to use the given holistic rubric, diversity in the scores assigned to translation papers could not be eliminated to a large extent. This suggests that the dissimilar ratings among raters might be related to how they understood and interpreted the given scale and to the criteria to which they gave priority.

Third, G-theory analysis indicated that with respect to the impact of the rater facet on holistic score variability in the three measurement designs, the raters displayed similar trends regarding leniency and severity when scoring two qualities of translation paper. However, the rater facet was determined to have no impact on the score variability, collectively. This revealed that the raters varied considerably in the scores they gave to high- and low-quality papers and they displayed great differences regarding severity and leniency within each translation paper quality. Additionally, in all three designs ($p \times r$ for high- and low-quality translation papers, and $p \times r \times q$ for all translation papers), the contributions of the residual facet were determined to be substantially large, indicating the impact of some unknown facets on the variability in scores. For example, in this research, text type or method of rating, which might have impact on score variability, were not taken into consideration.

Fourth, considering inter-rater reliability, high generalizability and dependability coefficients were yielded from the ratings of mixed-quality papers. Although the scorings of the high-quality papers yielded high generalizability coefficients, the dependability coefficients were relatively low. As for the low-quality papers, the scorings yielded low generalizability and dependability coefficients. This finding indicated that the raters were more consistent in the scores they gave to high-quality papers than to low-quality ones. These results might be related to the number of raters, given that when the number of facets are increased higher dependability coefficients are likely to be obtained. Also, the D-studies employed revealed that a lower number of raters would still yield acceptable dependability scores for mixed-quality translation papers. However, for low- and high-quality translation papers, acceptable dependability scores would be attained only when the number of the raters was increased.

Fifth, regarding gender impact on the reliability of scores, no significant differences were determined between female and male raters in terms of their ratings of high- and low-quality translation papers. However, although the difference was not significant between the male and female raters, male raters tended to assign higher scores to the translation papers collectively in comparison to the female raters. Regarding the impact of previous experience

on the variability and reliability of scores, it was determined that experienced raters assigned considerably lower scores to the translation papers compared to the inexperienced raters. Also, high-quality papers scored by inexperienced raters yielded higher reliability coefficients than those scored by more experienced raters. Similarly, regarding scorings of the low-quality papers, the reliability coefficients obtained from inexperienced raters were found to be higher than those of more experienced raters. Thus, the previous experience of raters was found to be a significant factor affecting the raters' scoring of translation papers.

Lastly, regarding the use of decision-making strategies, the instructors were found to apply significantly more judgment strategies than interpretation strategies while assigning scores to all translation papers. Also, they used significantly more strategies while grading low-quality translation papers when compared to high-quality translation papers. Regarding the use of strategies based on focus, the raters were found to prefer mostly strategies with a self-monitoring focus, followed by a language focus and a rhetorical/ideational focus, respectively. They followed the same trend for both high- and low-quality translation papers. In addition, the raters were observed to display different degrees of severity and leniency while scoring high- and low-quality translation papers, even though they had been given a thorough training in following the criteria in the holistic rubric.

Pedagogical Implications

In light of the limitations and results of this research, possible pedagogical implications are suggested. First, the main purpose of this study was not to compare or contrast expert and novice instructors, but it examined the differences between these two experience groups. Variations between the scorings of the raters indicated that inexperienced instructors were more consistent in the scores they gave to the papers; however, experienced raters were more inconsistent in their scorings and they might have their own criteria. Although training in scoring was provided to the raters before the procedure of assessment of the students' translation papers, the variations between the scorings of the raters were significant. Therefore, this study shows that even raters with broad experience of rating should be given detailed and consistent rater training to make them more reliable markers. In this way, variation between the scorings could be eliminated to a considerable extent and fairer judgments could be attained.

Secondly, it is suggested that in-house scoring protocols and thorough rater training might be helpful for instructors to achieve fairer judgments. Seven out of the ten raters in this study were from the ELL Department of Kafkas University. Although two translation assessment exams (mid-term and final) are held in each semester of an academic year, there is no available protocol for scoring translation performance assessment. Also, the program does

not provide a double-grading system. Thus, matching inexperienced and experienced raters in pairs to score the translation papers of students might yield more reliable ratings, as in this study experienced raters were found to be more severe than the inexperienced raters in their scorings. That is, when two experienced raters are matched for the scoring of a translation paper, they are likely to assign lower scores. However, when the scoring is carried out by two inexperienced raters, they might assign higher scores to the same paper. Thus, pairing inexperienced and experienced raters might substantially eliminate these consequences in double-grading circumstances.

Moreover, although holistic rubrics are considered to be efficient when the focus is on the overall quality of a performance, and are quick to use and provide the rater with a snapshot of the performance at hand (Bargainner, 2003), the results of this study revealed significant variation in the scorings of the raters. Thus, the use of an analytic rubric with more detailed descriptors might provide more specific results and might yield more reliable and consistent scorings (Charney, 1984; Elbow, 1999; Shi, 2001; Weigle, 2002). Also, the findings of this study showed that, instead of applying a traditional holistic or analytic rubric, developing a more user-friendly rubric with more criteria might be useful for eliminating rating variations between the raters.

Furthermore, in this study, students were given four texts belonging to a single genre (informative) to translate from English to Turkish. Thus, assessing students' translation performance over a single genre might be limited with regard to generalizing the scores since text type might be a factor impacting on score variability (Gebril, 2009; Weigle, 1999). In this regard, evaluating students' translation performance and the rating performance of raters through different text types might be helpful in minimizing such pedagogical concerns.

Lastly, the results of this research revealed that experience is a factor that has an impact on performance ratings. Also, previous research has shown that the demographics of raters, such as age (Cleveland & Landy, 1981; Griffeth & Bedeian, 1989), education (Cascio & Valenzi, 1977) and gender (Northcraft, Huber, & Neale, 1988); and some other characteristics, such as rater mood (Sinclair, 1988; Fried, Levi, BenDavid, Tiegs, & Avital, 2000), and rater attitude (Tziner, Murphy, & Cleveland, 2001) might impact scorings. Thus, further research investigating these factors might be helpful for understanding the scoring behaviors of raters.

Methodological Implications

Various methodological implications can be drawn from the results and limitations of this research. Firstly, examining the translation rating performances of instructors through only

a holistic rubric brings about a limitation in terms of the generalizability of the findings. Thus, a comparison of the scoring performances of instructors regarding variability of scores and rater severity involving the use of different types of rubric might yield more insight into translation evaluation in this research context (e.g. Han, 2013; Şahan, 2018; Weigle, 2002). Besides, investigating the decision-making behaviors of raters through various types of scales within the same context of study may present a deeper insight into the strategies applied by raters while assigning scores to papers.

Another methodological implication might be drawn about the context of the research. In order to create stress-free environments for both students and instructors, this research was conducted under flexible conditions. The students were given adequate time to complete translation tasks and they were allowed to use any type of dictionaries they wanted. Similarly, raters were free to take the translation papers to their homes and assign scores to them whenever they wanted. However, different results might be obtained in a more naturalistic context where students are asked to complete translation tasks under real-life exam conditions, and where raters are allocated a limited time span to assign scores to the papers under the same controlled conditions.

Another implication might be suggested regarding the use of thinking aloud protocols. In this research, the researcher scored one of the translation papers in order to show the raters how to conduct TAPs, and after the scoring process had been video-recorded, it was uploaded to YouTube. Then, the video was shared with the instructors. Also, in order to make sure that the raters understood the think aloud procedure fully, the researcher administered a TAP training process. However, the findings showed that raters mostly applied the criteria in their minds and two of the raters failed to complete the TAP tasks. Thus, it can be said that the theoretical training was not detailed or sufficient enough to make the raters fully understand how to conduct TAPs while rating translation papers. In this regard, conducting more detailed and thorough training might enhance the quality of the data obtained from verbal protocols (e.g. Şahan, 2018). For instance, to make the raters better verbalize what they think, they might be requested to rate a sample paper and, in the light of that experience, a follow-up session might be held to discuss unclear and problematic aspects of the procedure.

Further Research

Based on the limitations and results of this research, a number of recommendations for future research are given. Firstly, in this study, the use of holistic assessment was considered to be more appropriate since the translation tasks given to the students had no definite correct answer and the focus of scoring was on the overall quality of the translation papers. The holistic

rubric used in the study included six separate components to be scored individually and summed the sub-scores to gain a total score. Therefore, as in analytical assessments, the holistic scoring used in this research was expected to yield reliable and consistent scores. However, the raters who participated in the present study showed significant variation in the scores they gave to the translation papers of low-proficient students. Therefore, future studies could examine the impact of rater expectations and assessment standards upon the holistic and analytic rating of translation papers, since the raters in this study seemed to follow the criteria in their minds rather than conforming to the criteria in the given holistic rubric.

Secondly, the use of a holistic scoring approach is more desirable in situations where the assessment purpose is summative in nature (Mertler, 2000). Also, it is used to provide more flexibility to raters in considering different aspects of a written performance. The holistic rubric preferred in this study included various aspects on which raters based their judgments and assigned scores while assessing the translation papers. However, this choice brought some problems related to the reliability of the scorings particularly for low-quality translation papers. Instead of considering different aspects of the papers, the raters seemed to focus merely on the aspects given in the holistic scale. In addition, although holistic rubrics are considered to be more time saving and practical than analytic rubrics, they provide limited feedback regarding strengths and weaknesses in a written task (Mertler, 2000). For this reason, the use of an analytic rubric with more detailed traits and aspects might give a deeper understanding of the decision-making behaviors of instructors. Therefore, future research might involve instructors in the process of developing a scoring rubric which contains both analytic and holistic approach features, rather than using an existing scale.

Thirdly, the findings of this study revealed that experience is an important parameter impacting the variability and reliability of translation scorings. In addition, it was found to have great impact on rater severity and leniency. In the present study, experienced raters were found to be more severe and more inconsistent in the scorings they assigned to both qualities of translation papers when compared to novice raters. Thus, the impact of rater experience on score variability could be examined in further studies using both quantitative and qualitative research methods. Determining the reasons behind experienced and inexperienced raters' scoring attitudes might shed light on the issues of reliability and variability in scoring in translation or writing assessment.

Fourthly, in this study, texts translated from English to Turkish were used for data collection since the students who participated in this study were taking translation courses from English to Turkish during the course of the study. However, translation is a two-way activity

and text type might have impact on learners' translation performance and raters' rating behaviors. Thus, further research might examine the impact of texts translated from Turkish to English on variability of scores in translation assessment.

Lastly, although higher education policy is regulated by central authorities in Turkey, universities determine their own assessment procedures and practices for evaluating student performance. Thus, the employment of different grading procedures at different institutions and even within the same institution raises a fairness problem in Turkish universities. In this study, the impact of rater training and practices of translation assessment on the reliability of scorings was examined to provide specific areas for consideration in relation to large-scale national and institutional assessments. However, rather than conducting a traditional rater-training procedure centering upon the use of a particular scoring scale, further research might implement a strategy-based training program to better understand the decision-making behaviors of raters and to address the requirements of specific writing assessment contexts.

REFERENCES

- Aksoy, B. (2001). Çeviride eleştirmen seçimleri ışığında çeviri eleştirisi. *Hacettepe Üniversitesi, Edebiyat Fakültesi Dergisi*, 18(2), 1-16.
- Aktaş, T. (1996). *Çeviri işlemine genel bir bakış*. Ankara: Orsen Matbaacılık.
- Al-Melhi, A. (2014). Pedagogical implications of utilizing translation evaluation strategies with translation students: Toward a model of teaching translation. *Arab World English Journal*, 3, 11-26.
- Al-Qinai, J. (2000). Translation quality assessment: Strategies, parameters and procedures. *Meta: Translators' Journal*, 45(3), 497-519. <https://doi.org/10.7202/001878ar>
- Ammann, M. (1990). Anmerkungen zu einer Theorie der Übersetzungskritik und ihrer praktischen Anwendung. *TexTconText*, 5, 209-250.
- Anckaert, Ph., Eyckmans, J., Justens, D. & Segers, W. (2013). Bon sens, faux-sens, contresens et non-sens sens dessus dessous: Pour une évaluation fidèle et valide de la compétence de traduction. In J.Y. Le Dizez & W. Segers (Eds.), *Le bon sens en traduction* (pp. 79–94). Rennes: Presses Universitaires de Rennes.
- Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13-19.
- Angelelli, C. V. (2009). *Testing and assessment in translation and interpreting studies*. Philadelphia: John Benjamins.
- Angelelli, C. V. & Jacobson, H. (2009). *Testing and assessment in translation and interpreting studies: A call for dialogue between research and practice*. Amsterdam: John Benjamins Publishing Company.
- Ateşman, E. (2001). Kültürel farklardan kaynaklanan çeviri sorunları ve çözüm önerileri. *Hacettepe Üniversitesi Edebiyat Fakültesi Dergisi*, 18(2), 29-35.
- Aybirdi, N. (2019, July 18). A sample think-aloud protocol [Video File]. Retrieved from <https://www.youtube.com/channel/UCb5-0KcxHMuRo4FLO9d26aw>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. (2010). *Language testing in practice*. Oxford: Oxford University Press.
- 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.
- Baker, A. B. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, 9(3), 225-248. <https://doi.org/10.1080/15434303.2011.637262>
- Bargainnier, S. (2003). Fundamentals of rubrics. Retrieved from <http://www.pacific-crest.com>
- Barghout, M. A. M. (1990). *Translation quality assessment: An application of a rhetorical model*. Doctoral dissertation, University of Salford, UK.
- Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished doctoral dissertation, University of Toronto, Canada.

- 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>
- Basnett, S. (1991). *Translation studies (Revised Edition)*. London/New York: Routledge.
- Berggren, I. (1972). *Does the use of translation exercises have negative effects on the learning of a second language?* Gothenburg: Department of English, Gothenburg University.
- Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, 18(1), 57-75. <https://doi.org/10.1080/0729436990180105>
- Biggs, J. (2003). *Aligning teaching for constructing learning*. York, UK: Higher Education Academy.
- Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, 1(5), 5-22.
- Birel, A. S., & Albuz, A. (2014). Viyolonsel öğretiminde performansı değerlendirmeye yönelik hazırlanan dereceli puanlama anahtarının (rubrik) sınanması ve değerlendirilmesi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 18(3), 181-207.
- Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill Book Company.
- Boztaş, İ. (1992). *Çeviri, çeviride eşdeğerlilik ve dilbilim*. Ankara: Karaca Dil Kursu Yayınları.
- Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review*. Princeton, NJ: Educational Testing Service.
- Brennan, R. L. (2001). *Generalizability theory*. New York, NY: 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>
- Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology*, 52(1), 13-15. <https://doi.org/10.1016/j.jsp.2013.11.008>
- Brown, G. (2001). *Assessment: A guide for lecturers* (Vol. 3). LTSN Generic Centre, York: LTSN.
- Brown, H. D. (2004). *Language assessment: Principles and classroom practice*. New York, NY: Pearson/Longman.
- 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>
- Carreres, Á., & Noriega-Sánchez, M. (2011). Translation in language teaching: Insights from professional translator training. *The Language Learning Journal*, 39(3), 281-297. <https://doi.org/10.1080/09571736.2011.567356>
- Carroll, J. B. (1966). An experiment in evaluating the quality of translations. *Mechanical Translation and Computational Linguistics*, 9(3), 55-66.

- Carroll, J. B. (1991). Response to Don Porter's paper: "Affective factors in language testing." In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 41–45). London: Modern English Publications/British Council.
- Cary, E., & Jumpelt, R. W. (1963). *Quality in translation*. Oxford/London/New York/Paris: Pergamon Press.
- Cascio, W. F., & Valenzi, E. R. (1977). Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology*, 62(3), 278-282. <https://doi.org/10.1037/0021-9010.62.3.278>
- Catford, J. C. (1965). *A linguistic theory of translation; An essay in applied linguistics*. London: Oxford University Press
- Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English*, 18(1), 65-81.
- Cleveland, J. N., & Landy, F. J. (1981). The influence of rater and ratee age on two performance judgments. *Personnel Psychology*, 34(1), 19-29. <https://doi.org/10.1111/j.1744-6570.1981.tb02173.x>
- Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle Publishers.
- Cohen, A. D., & Brooks-Carson, A. (2001). Research on direct versus translated writing: students' strategies and their results. *The Modern Language Journal*, 85(2), 169-188. <https://doi.org/10.1111/0026-7902.00103>
- 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>
- Colina, S. (2015). *Fundamentals of translation*. Cambridge: Cambridge University Press.
- Conde, T. (2011). Translation evaluation on the surface of texts: a preliminary analysis. *The Journal of Specialized Translation*, 15, 69-86.
- Connor-Linton, J. (1995). Looking behind the curtain: What do L2 compositions really mean? *TESOL Quarterly*, 29(4), 762-765. <https://doi.org/10.2307/3588174>
- Cook, G. (2010). *Translation in language teaching: An argument for reassessment*. Oxford: Oxford University Press.
- Cresswell, J. W. (2011). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). New Delhi, India: PHI Learning Private.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks: Sage Publications.
- Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Thousand Oaks, CA: Sage.
- Cronbach, L. J., Gleser G. C., and Nanda H. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley.
- Cuellar, S. B. (2002). Reseña de " Translation quality assessment. A model revisited." de House Juliane. *Forma y Función*, 15, 181-193.
- Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing*, 7(1), 31-51. <https://doi.org/10.1177/026553229000700104>

- Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series, Report No: 22). Princeton, NJ: Educational Testing Service.
- 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>
- Darbelnet, J. (1970). Traduction littérale ou traduction libre? *Meta*, 15(2), 88-94. <https://doi.org/10.7202/002478ar>
- De Gruijter, D. N. (1984). Two simple models for rater effects. *Applied Psychological Measurement*, 8(2), 213-218. <https://doi.org/10.1177/014662168400800210>
- De Gruijter, D. N., & Leo, J. T. (1991). Generalizability theory. In R. K. Hambleton, & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 45-68). Springer, Dordrecht.
- Dolan, M. G. (2005). Reliability: current issues and concerns. *Athletic Therapy Today*, 35-39. Retrieved from <https://pdfs.semanticscholar.org/bb4b/b8522ec7b33cdb8b569b414c268b9aa6b8d7.pdf>
- Drugan, J. (2013). *Quality in professional translation: Assessment and improvement*. UK & USA: Bloomsbury Academic.
- Ebbert-Hübner, C. & Maas, C. (2017). Can translation improve EFL students' grammatical accuracy? *International Journal of English Language & Translation Studies*, 5(4), 191-202.
- 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. Retrieved from ERIC: <https://files.eric.ed.gov/fulltext/ED374661.pdf>
- Elbow, P. (1999). Ranking, evaluating, and liking: Sorting out three forms of judgements. In R. Straub (Ed.), *A sourcebook for responding to students writing* (pp. 175-196). Cresskill, NJ: Hampton Press.
- Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly*, 1(1), 2-24.
- Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing writing*, 18(2), 111-131. <http://dx.doi.org/10.1016/j.asw.2012.12.002>
- Eyckmans, J. & Anckaert, P. (2017). Item-based assessment of translation competence: Chimera of objectivity versus prospect of reliable measurement. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 16, 40-56.
- 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). Amsterdam: John Benjamins Publishing Company.
- Faiq, S. (2005). Cultural dislocation through translation. *Intercultural Communication Studies*, 14(4), 57-75.

- 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.
- Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English*, 15(3), 245-255.
- Fried, Y., Levi, A. S., Ben- David, H. A., Tiegs, R. B., & Avital, N. (2000). Short research note: Rater positive and negative mood predispositions as predictors of performance ratings of ratees in simulated and real organizational settings: Evidence from US and Israeli samples. *Journal of Occupational and Organizational Psychology*, 73(3), 373-378.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London and New York, NY: Routledge.
- Fumani, M. R. F. Q. (2013). An assessment of the translations of English lyrics in World's Six Different Music Style Based on Hurtado's Model of error analysis. *Linguistics and Literature Studies*, 1(3), 157-163.
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all?. *Language Testing*, 26(4), 507-531. <https://doi.org/10.1177/0265532209340188>
- Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15(2), 100-117. <https://doi.org/10.1016/j.asw.2010.05.002>
- Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, 21(2), 56-73. <http://dx.doi.org/10.1016/j.asw.2014.03.002>
- 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.
- Ghonsooly, B. (1993): Development and validation of a translation test. *Edinburgh Working Papers in Applied Linguistics*, 4, 54-62.
- Goodwin, L. D., & Goodwin, W. L. (1991). Using generalizability theory in early childhood special education. *Journal of Early Intervention*, 15(2), 193-204. <https://doi.org/10.1177/105381519101500208>
- Gouadec, D. (2007). *Translation as a profession*. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Griffeth, R.W. & Bedeian, A.G. (1989). Employee performance evaluations: Effects of ratee age, rater age, and ratee gender. *Journal of Organizational Behavior*, 10, 83-90. <https://doi.org/10.1002/job.4030100107>
- Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Education and Science*, 34(154), 93-103.
- Güler, N., Eroğlu, Y., & Akbaba, S. (2014). Reliability of criterion-dependent measurement tools according to generalizability theory: application in the case of eating skills. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 14(2), 217-232. <https://doi.org/10.17240/aibuefd.2014.14.2-5000091536>
- Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi Yayınları.

- Gündüz, N. (2016). The scope and use of ESP in foreign and Turkish context. *Türkiye Sosyal Araştırmalar Dergisi*, 20(3), 647-669.
- Hall, C. B. (2015). Comment: Generalizability theory and assessment in medical training. *Neurology*, 85(18), 1628-1628.
- Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An investigation of problems and solutions*. Unpublished doctoral dissertation, Atatürk University, Turkey.
- Han, T. (2017). Scores assigned by inexperienced raters to different quality of EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education*, 13(1), 136-152.
- Harrison, R. L. (2013). Using mixed methods designs in the Journal of Business Research, 1990–2010. *Journal of Business Research*, 66(11), 2153-2162. <https://doi.org/10.1016/j.jbusres.2012.01.006>
- Hatim, B., & Mason, I. (1997). *The translator as communicator*. New York, NY: Routledge
- Heitman, R. J., Kovaleski, J. E., & Pugh, S. F. (2009). Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. *Journal of Athletic Training*, 44(1), 48-52. <https://doi.org/10.4085/1062-6050-44.1.48>
- Hönig, H. (1998). Positions, power and practice: Functionalist approaches and translation quality assessment. In C. Schäffner (Ed.) *Translation and quality*, (pp. 6-34). Clevedon: Multilingual Matters.
- House, J. (1977). *A modal for translation quality assessment*. Tübingen: TBC Verlag Gunter Narr.
- House, J. (1981). *A model for translation quality assessment*. Tübingen: Gunter Narr.
- House, J. (1997). *Translation quality assessment: A model revisited*. Tübingen: Gunter Narr.
- House, J. (1998). *Translation quality assessment: A model revisited*. Tübingen: Gunter Narr.
- House, J. (2001). Translation quality assessment: Linguistic description versus social evaluation. *Meta: Translators' Journal*, 46(2), 243-257. <https://doi.org/10.7202/003141ar>
- House, J. (2014). *Translation quality assessment past and present*. New York: Routledge.
- House, J. (2015). *Translation quality assessment past and present*. New York: Routledge.
- 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. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal*, 2(4), 423-443. <https://doi.org/10.5054/tj.2011.269751>
- Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large scale ESL writing assessment. *Assessing Writing*, 17(3), 123-139. <https://doi.org/10.1016/j.asw.2011.12.003>
- Huang, J., & Foote, C.J. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, 7(3), 219-233. <https://doi.org/10.1080/15434300903540894>

- 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). New York, NY: Untested Ideas Research Center.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press
- Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, (pp. 206-236). Cresskill, NJ: Hampton Press.
- Hurtado Albir, A. (2015). The acquisition of translation competence. Competences, tasks, and assessment in translator training. *Meta: Journal des Traducteurs/Meta: Translators' Journal*, 60(2), 256-280. <https://doi.org/10.7202/1032857ar>
- Ibrahim, A. M. (2011). Using generalizability theory to estimate the relative effect of class size and number of items on the dependability of student ratings of instruction. *Psychological Reports*, 109(1), 252-258. <https://doi.org/10.2466/03.07.11.PR0.109.4.252-258>
- 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>
- Jamshidian, E., & Mohammadi, R. (2012). Translation analysis and assessment of poetic discourse: Extra-textual meaning in Persian and English. *The Journal of International Social Research*, 5(21), 1-10.
- Jiménez-Crespo, M. A. (2011). A corpus-based error typology: Towards a more objective approach to measuring quality in localization. *Perspectives, Studies in Translatology*, 19(4), 315-338. <https://doi.org/10.1080/0907676X.2011.615409>
- Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks*. New York: Guilford.
- 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). Manchester: 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). New York: Routledge.
- Kelly, L. G. (1979). *The true interpreter: History of translation theory and practice in the West*. Oxford: Basil Blackwell.
- Kelly, N. & Bruen, J. (2015). Translation as a pedagogical tool in the foreign language classroom: A qualitative study of attitudes and behaviors. *Language Teaching Research*, 19(2), 150-168. <http://dx.doi.org/10.1177/1362168814541720>
- Khanmohammad, H., & Osanloo, M. (2009). Moving toward objective scoring: A rubric for translation assessment. *Journal of English Language Studies*, 1(1), 131-153.
- Kharmandar, M. A. (2017, November). *A Critique of Argumentation-Centered Translation Quality Assessment: Beyond Toulmin's Model*. Paper presented at the First National Conference on Innovative Multidisciplinary Research in the Humanities. Retrieved from https://s3.amazonaws.com/academia.edu.documents/51388539/A_Critique_of_Argumentation-Centered_Translation_Quality_Assessment.pdf?

- Kieffer, K. M. (1998). "Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate?" [Proceeding]. Paper Presented at the Annual Meeting of the South Western Psychological Association. New Orleans, LA, USA.
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Flesch-Kincaid grade level*. Memphis: United States Navy.
- Kis B., & Mohácsi-Gorove A. (2008). *A fordító számítógépe*. Bicske: Szak Kiadó.
- Kobayashi, H., & Rinnert, C. (1992). Effects of first language on second language writing: translation versus direct composition. *Language Learning*, 42(2), 183- 215. <https://doi.org/10.1111/j.1467-1770.1992.tb00707.x>
- Köksal, D. (2005). *Çeviri eğitimi: Kuram ve uygulama*. Ankara: Nobel Yayın Dağıtım.
- Kraiger, K., & Teachout, M. S. (1991). *Application of Generalizability Theory to the Air Force Job Performance Measurement Project: A summary of research results*. Colorado: Univ At Denver Dept Of Psychology.
- Lado, R. (1964). *Language teaching, a scientific approach*. New York: McGraw-Hill.
- Lakes, K. D. (2012). The Response to Challenge Scale (RCS): The development and construct validity of an observer-rated measure of children's self-regulation. *The International Journal of Educational and Psychological Assessment*, 10(1), 83-96.
- Lakes, K. D. (2013). Restricted sample variance reduces generalizability. *Psychological Assessment*, 25(2), 643-650. <https://doi.org/10.1037/a0030912>
- Lakes, K. D., & Hoyt, W. T. (2009). Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research. *Journal of Clinical Child & Adolescent Psychology*, 38(1), 144-165. <https://doi.org/10.1080/15374410802575461>
- Lakes, K. D., Guo, Y., Lucas, C. T., & Cooper, D. (2017). Measuring maternal behaviors in the neonatal intensive care unit. *Infants and Young Children*, 30(2), 124-132. <https://doi.org/10.1097/IYC.0000000000000091>
- Lambert, J. (1991). Shifts, oppositions and goals in translation studies: towards a genealogy of concepts. In K. van Leuven-Zwart & T. Naaijken (Eds.), *Translation studies: The state of the art*, (pp. 25-37). Netherlands: Rodopi.
- 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>
- Lauscher, 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>
- Laviosa, S. (2014). *Translation and language education: Pedagogical approaches explored*. Routledge: New York.
- Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, 12(3), 237-255. https://doi.org/10.1207/S15324818AME1203_2
- 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). UK: Cambridge Scholars Publishing.

- Lei, P. W., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single- subject observational research. *Psychology in the Schools*, 44(5), 433-439. <https://doi.org/10.1002/pits.20235>
- Leuven-Zwart, K. (1990). Translation and original. similarities and dissimilarities II. *Target*, 2(1), 69-95.
- Levefere, A. (1996). Übersetzen, Übersetzung und Macht. In S. Heidemarie (Ed.) *Dolmetscher- und übersetzer- ausbildung gestern, heute und morgen, akten des internationalen wissenschaftlichen kolloquiums anlässlich des 100 jährigen jubiläums der dolmetscher- und übersetzer- ausbildung Russisch an der Berliner Universität*, (pp. 171-180). Frankfurt: Lang.
- Li, G., Hou, G., Wang, X., Yang, D., Jian, H., & Wang, W. (2018). A Multivariate Generalizability Theory Approach to College Students' Evaluation of Teaching. *Frontiers in Psychology*, 9, 1065. <https://doi.org/10.3389/fpsyg.2018.01065>
- Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, 86(2), 255. <https://doi.org/10.1037/0021-9010.86.2.255>
- Littlefield, J. H., Harrington, J. T., Anthracite, N. E., & Garman, R. E. (1981). A description and four-year analysis of a clinical clerkship evaluation system. *Journal of Medical Education*, 56(4), 334-340. <https://doi.org/10.1097/00001888-198104000-00006>
- Lo, S. W. (2016). *Using translation in L2 classrooms: An empirical study on non-language major students' engagement in class discussions and improvement in language usage*. Doctoral dissertation, School of Modern Languages, University of Leicester, England.
- Lumley, T. (2000). *The process of the assessment of writing performance: The rater's perspective*. Unpublished doctoral dissertation, University of Melbourne, Australia.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276. <https://doi.org/10.1191/0265532202lt230oa>
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3
- Machida, S. (2008). A step forward to using translation to teach a foreign/second language. *Electronic Journal of Foreign Language Teaching*, 5(1), 140-155. Retrieved from <http://e-flt.nus.edu.sg/v5sp12008/machida.pdf>
- Marais, K. (2013). Constructive alignment in translator education: reconsidering assessment for both industry and academy. *Certification*, 5(1), 13-31.
- Martinez Mateo, R. (2016). Aligning qualitative and quantitative approaches in professional translation quality assessment. *Encuentro*, 25, 36-44.
- Martínez Mateo, R., Montero Martínez, S., & Moya Guijarro, A. J. (2017). The modular assessment pack: A new approach to translation quality assessment at the Directorate General for Translation. *Perspectives*, 25(1), 18-48. <https://doi.org/10.1080/0907676X.2016.1167923>
- McAlester, G. (2000). The evaluation of translation into a foreign language. In B. Schaffner & B. Adab, *Developing translation competence*, (pp.229-241). Amsterdam: Benjamins.
- McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

- McNamara, T. F. (1996). *Measuring second language performance*. London and New York, NY: Addison Wesley Longman.
- Medadian, G. ve Mahabadi, D. N. (2015). A summative translation quality assessment model for undergraduate student translations: Objectivity versus manageability. *Studies about Languages*, 26, 40-54.
- Melis, N. M. ve Albir, A. H. (2001). Assessment in translation studies: Research needs. *Meta*, 46(2), 272-287.
- Mendelsohn, D., & Cumming, A. (1987). Professor's ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal*, 5(1), 09-26. <https://doi.org/10.18806/tesl.v5i1.512>
- Mertler, C. A. (2000). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation*, 7(25), 1-9. <https://doi.org/10.7275/gcy8-0w24>
- Mohácsi-Góove, A. (2014). *A minőség fogalma a fordítástudományban és a lektorálás mint minőségbiztosítási garancia*. Doktori disszertáció, Budapest, ELTE.
- Mohácsi-Gorove, A. (2014). The concept of translation quality in translation studies. Revision or reviewing as a guarantee of quality assurance. Retrieved from <http://doktori.btk.elte.hu/lingv/mohacsigoroveanna/thesis.pdf>
- Mohaghegh, A. (2011). Clarke's (1891) translation of Ghazal 167 by Hafez: A text-based approach in criticism at linguistic level based on Reiss'(1971) model. *International Social Research*, 4(18), 76-85.
- Mossop, B. 2001. *Revising and Editing for Translators*. Manchester: St. Jerome Publishing.
- Munday, J. (2001). *Introducing translation studies*. London and New York: Routledge
- Munday, J. (2016). *Introducing translation studies: Theories and applications*. Newyork: Routledge.
- Nerudová, L. (2012). *Quality of translation: approaches and a field survey*. Unpublished Doctoral dissertation, Masarykova Univerzita, Czechia.
- Neves, R. R. (2002). Translation quality assessment for research purposes: An empirical approach. *Cadernos de Tradução*, 2(10), 113-131.
- Newmark, P. (1991). *About translation*. Clevedon: Multilingual Matters.
- Newmark, P. (1998). *A textbook of translation*. London: Prentice Hall
- Nida, E. (1964). *Towards a science of translating with special reference to principles and procedures involved in Bible translating*. Brill: Leiden.
- Nida, E. A. (1964). *Toward a science of translating*. Brill: Leiden.
- Nida, E. A. (1982). *The theory and practice of translation*. Leiden: Brill.
- Nida, E. A., & Taber, Ch. R. (1969). *The theory and practice of translation*. Brill: Leiden.
- Nightingale, S., Carew, A., & Fung, J. (2007). *Application of constructive alignment principles to engineering education: Have we really changed?* Paper presented Proc. of AaeE Conference, Melbourne, Australia.
- Nobar, N. & Navidpour, P. (2010). Translating poetry: Based on textual and extra textual analysis. Retrieved on October 19, 2019 from <http://www.translationdirectory.com/articles/article2125>.

- Nord, C. (1992). Text analysis in translator training. In: C. Dollerup, & A. Loddegaard (Eds.), *Teaching translation and interpreting*, (pp. 39-48). Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Nord, C. (1997). Defining translation functions. The translation brief as a guideline for the trainee translation. *Ilha do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, 33, 39-54. <https://doi.org/10.5007/%25x>
- Nord, C. (2007). Function plus loyalty: Ethics in professional translation. *Génesis*, 6, 1-14.
- Northcraft, G. B., Huber, V., & Neale, M. A. (1988). Sex effects in performance--related judgments. *Human Performance*, 1(3), 161-175.
- Oosterveld, P., & Cate, O. T. (2004). Generalizability of a study sample assessment procedure for entrance selection for medical school. *Medical Teacher*, 26(7), 635-639. <https://doi.org/10.1080/01421590400004874>
- Orozco, M. (2000). Building a measuring instrument for the acquisition of translation competence in trainee translators. *Benjamins Translation Library*, 38, 199-214.
- Pöschhacker, F. (1994) *Simultan dolmetschen als komplexes handeln*. Tübingen: 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/10.2307/329727>
- Prion, S. K., Gilbert, G. E., & Haerling, K. A. (2016). Generalizability theory: an introduction with application to simulation evaluation. *Clinical Simulation in Nursing*, 12(12), 546-554. <https://doi.org/10.1016/j.ecns.2016.08.006>
- Prunč, E. (2007). Priests, princes and pariahs: Constructing the professional field of translation. In M. Wolf & A. Fukari (Eds.), *Constructing a sociology of translation*, (pp. 39-56). Amsterdam, Netherlands: John Benjamins Publishing Company.
- Pym, A. (2010). *Exploring translation theories*. New York: Routledge.
- Rabadán, R., Labrador, B., & Ramón, N. (2009). Corpus-based contrastive analysis and translation universals: A tool for translation quality assessment English-->Spanish. *Babel*, 55(4), 303-328. <https://doi.org/10.1075/babel.55.4.01rab>
- Radzi, N. M. M., Nor, M. M., & Matore, M. E. E. M. (2018). Using generalizability theory in the development of performance assessment of the cognitive domain of children. *Advanced Science Letters*, 24(7), 5297-5301.
- Reiss, K. & Vermeer, H. (1984). *Grundlegung einer allgemeinen translations theorie*. Tübingen: Niemeyer.
- Reiss, K. (1971) *Möglichkeiten und grenzen der übersetzungskritik. Kategorien und kriterien für eine sachgerechte beurteilung von übersetzungen*. München: Max Hueber Verlag
- Reiss, K. (1983). *Texttyp und übersetzungsmethode: Der operative text*. Heidelberg: Groos.
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.
- 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>
- Şahan, Ö. (2018). *The impact of rating experience and essay quality on rater behavior and scoring*. Unpublished doctoral dissertation, Çanakkale Onsekiz Mart University, Turkey

- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493. <https://doi.org/10.1177/0265532208094273>
- Scherer, G. A. C. & Wertheimer, M. (1964). *A psycholinguistic experiment in foreign-language teaching*. New York: McGraw-Hill.
- Schjoldager, A. (2003). Translation for language purpose: Preliminary results of an experimental study of translation and picture verbalization. *HERMES-Journal of Language and Communication in Business*, 30, 199-213. <https://doi.org/10.7146/hjlc.v16i30.25722>
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22(1), 1-30. <https://doi.org/10.1191/0265532205lt295oa>
- Schultz, K., & Whitney, D. (2005). *Classical true score theory and reliability*. Thousand Oaks, CA: Sage.
- Scriven, M. (1967). *The methodology of evaluation*. Chicago: Rand McNally and Co.
- Sezer, S. (2005). Öğrencinin akademik başarısının belirlenmesinde tamamlayıcı değerlendirme aracı olarak rubrik kullanımı üzerinde bir araştırma. *Pamukkale Üniversitesi, Eğitim Fakültesi Dergisi*, 18(18), 61-69.
- Shahraki, A., & Karimnia, A. (2011). Waddington's model of translation quality assessment: a critical inquiry. *Linguistics and Translation*, 40, 5219-5224.
- Sharkas, H. (2009). Translation quality assessment of popular science articles: Corpus study of the scientific American and its Arabic version. *Trans-kom*, 2(1), 42-62.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A premier*. Newbury Park, CA: Sage.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215-232.
- 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>
- Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18(3), 303-325. <https://doi.org/10.1177/026553220101800303>
- 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>
- Sinclair, R. C. (1988). Mood, categorization breadth, and performance appraisal: The effects of order of information acquisition and affective state on halo, accuracy, information retrieval, and evaluations. *Organizational Behavior and Human Decision Processes*, 42(1), 22-46. [https://doi.org/10.1016/0749-5978\(88\)90018-0](https://doi.org/10.1016/0749-5978(88)90018-0)
- Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating guidelines for test adaptations: A methodological analysis of translation quality. *Journal of Cross-Cultural Psychology*, 37(5), 557-567. <https://doi.org/10.1177/0022022106290478>
- 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>
- Soler, H. H. (2000). The effect of gender and working place of raters on university entrance examination scores. *Revista Española de Lingüística Aplicada*, 14, 161-180.

- 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](https://doi.org/10.1016/S1060-3743(96)90023-5)
- Sönmez, S. (1997). Çeviribilimin yabancı dil öğretiminde yeri ve önemi. In T. Kurultay, & İ. Birkandan (Eds.), *Forum: Türkiye’de çeviri eğitimi*, (pp. 31-39). İstanbul: Sel Yayıncılık.
- Speck, B. W. (2009). *Grading student writing: an annotated bibliography*. Westport, Conn: Greenwood Press.
- Stansfield, C. W., Scott, M. L., & Kenyon, D. M. (1992). The measurement of translation ability. *The Modern Language Journal*, 76(4), 455-467. <https://doi.org/10.2307/330046>
- Steiner, G. (1994). *After Babel. Aspects of language and translation*. Oxford: OUP.
- Stejskal, J. (2009). Quality assessment in translation. In M. Forstner, H. Lee-Jahnke & P. A. Schmitt (Eds.), *CIUTI-Forum 2008: Enhancing translation quality: ways, means, methods*, (pp. 291-300). Cambridge, Massachusetts: The MIT Press.
- Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, 59(3), 492-506.
- Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal*, 74(5), 49-55.
- Tavakol, M., & Brennan, R. L. (2013). Medical education assessment: a brief overview of concepts in generalizability theory. *International Journal of Medical Education*, 4, 221-222. <https://doi.org/10.5116/ijme.5278.a850>
- 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. <http://dx.doi.org/10.7575/aiac.ijels.v.5n.2p.47>
- Tindal, G., Yovanoff, P., & Geller, J. P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *The Journal of Special Education*, 44(1), 3-17.
- Toury, G. (1995). *Descriptive translation studies – and beyond*. Amsterdam and Philadelphia, PA: John Benjamin’s.
- Trejo-Mejía, J. A., Sánchez-Mendiola, M., Méndez-Ramírez, I., & Martínez-González, A. (2016). Reliability analysis of the objective structured clinical examination using generalizability theory. *Medical Education Online*, 21(1), 31-65. <http://dx.doi.org/10.3402/meo.v21.31650>
- Tsutsui, E., Owada, K., Kondo, Y., & Nakano, M. (2006). A diagnosis of Japanese EFL learners’ individual differences. In *Proceedings of the 11th Conference of Pan-Pacific Association of Applied Linguistics, Korea* (Vol. 11, pp. 281-289).
- Tziner, A., Murphy, K. R., Cleveland, J. N., & Roberts- Thompson, G. P. (2001). Relationships between attitudes toward organizations and performance appraisal systems and rating behavior. *International Journal of Selection and Assessment*, 9(3), 226-239.
- 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](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.
- VahidDastjerdi, H., Hakimshafaaii, H., & Jannesaari, Z. (2008). Translation of poetry: towards a practical model for translation analysis and assessment of poetic discourse. *Journal of Language and Translation*, 9(1), 7-40.
- Van Someren, M. W., Barnard, Y. F. ve Sandberg, J. A. C. (1994). *The think aloud method: A practical approach to modelling cognitive*. London: Academic Press
- Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In. L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 11-126). Norwood, NJ: Ablex.
- Vehmas-Lehto, I. (2010). Termit kääntäjän näkökulmasta. Käännösteoria, ammattikielet ja monikielisyys. *N Julkaisut*, 37, 361-372.
- Venuti, L. (1995) *The translator's invisibility. A history of translation*. London & New York: Routledge.
- Vermeer, H. J. (1978). Ein rahmen für eine allgemeine translationstheorie. *Lebende Sprachen*, 23(3), 99-102.
- Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, 23(1), 1-26. <https://doi.org/10.1037/met0000107>
- Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta*, 46(2), 331–325. <https://doi.org/10.7202/004583ar>
- 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). New Jersey: John 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](https://doi.org/10.1016/S1075-2935(00)00010-6)
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Williams, M. (2004). *Translation quality assessment: An argumentation-centered approach*. Ottawa: University of Ottawa Press.
- Williams, M. (2009). Translation quality assessment. *Mutatis Mutandis*, 2(1), 3–23.
- Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465-492. <https://doi.org/10.1177/0741088398015004002>
- Yamini, H., & Abdi, F. (2010). The application of House's Model on William Shakespeare's "Macbeth" and its Persian translation by Ala'uddin Pasargadi. Retrieved from https://www.translationdirectory.com/pdf/house_model.pdf
- Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education*, 21(3), 273-291. <https://doi.org/10.1080/08957340802161840>

APPENDICES

Appendix-A. Checklist for Holistic Rubric

1. Purpose 1.1.Determine/negotiate the viability of the brief. 1.2.Clarify the function/goal/purpose of the translation. 1.3.Determine the motive for the translation. 1.4.Formulate an overall translation strategy, which would help to create a text that would fulfill its purpose.
2. Culture 2.1.Analyze the requirements of the target culture. 2.2.Analyze the requirements of the target audience. 2.3.Analyze the time of reception. 2.4.Analyze the place of reception.
3. Text 3.1.Determine appropriate text type. 3.2.Determine appropriate subject field. 3.3.Determine appropriate style. 3.4.Determine appropriate register.
4. Visual Appearance 4.1.Consider visual appearance in terms of brief and target audience/culture. 4.2.Edit for formal layout, consistency, readability, etc. 4.3.Proofread.
5. Content 5.1.In terms of the brief and the analyses done under 1, ascertain whether information is accurate/adequate. 5.2.In terms of decisions taken under 1, decide what information from the source text to utilize in target text.
6. Language 6.1.Revise translation in terms of brief analysis. 6.2.Revise translation in terms of source text analysis. 6.3.Revise translation in terms of translation strategy. 6.4.Edit for consistency in language and idiom. 6.5.Edit for consistency in syntax. 6.6.Do research on issues of terminology.

Appendix-B. Rater Profile Form

This part of the study aims to gather background information related to your evaluations to understand your performance but not to judge your performance. Please note that your information and identity will remain confidential.

Your Name is.....

1. Your gender: *Male..... Female.....*

2. Your age: *20-30..... 31-40..... 41-50..... 50 or more.....*

3. What is your highest level of education?

B.A..... M.A..... PhD..... Other, please specify.....

4. How long have you been teaching EFL?

Less than 2 years..... 3 to 4..... 5 to 6..... 7 to 10..... More than 10 years.....

5. How long have you been teaching *EFL* writing?

Less than 2 years..... 3 to 4..... 5 to 6..... 7 to 10..... More than 10 years.....

6. What is your experience of marking EFL papers (both in classroom and large- scale assessment contexts)?

In classroom assessment,

I have experience ...

Less than 2 years..... 3 to 4..... 5 to 6..... 7 to 10..... More than 10 years.....

In large-scale assessment,

I am... *Less than 2 years..... 3 to 4..... 5 to 6..... 7 to 10..... More than 10 years.....*

7. Have had any training courses in assessing language performance before?

Yes..... No.....

8) What is your method frequently used for marking EFL papers?

Holistic..... Analytic.....

Thank you for your cooperation.

Appendix-C. Consent Form for Students

Written Consent Form (for students)

I, the undersigned, have received explanation from the researcher regarding the background, purpose, significance, and procedures of this research. The researcher has assured that my identity will remain confidential and I can freely withdraw my participation should I choose to. I have been given opportunities to ask questions, and the researcher has responded to them satisfactorily. With this form, I confirm that I am willingly and voluntarily participate in this research with no coercion or undue influences from any parties.

Name of participant

Signature of participant

Name of researcher

Signature of researcher

Date: _____

Appendix-D. Holistic Scoring Rubric

1. Purpose	
1.1 The translation fulfils the purpose for which it was created.	1
<i>Total</i>	1
2. Culture	
2.1 The translation is adapted to the requirements of the target culture, given the brief.	1
2.2 The translation is adapted to the requirements of the target audience, given the brief.	1
<i>Total</i>	2
3. Text	
3.1 The text type fulfils the requirements of the brief.	0.5
3.2 The target text is structured cohesively.	0.5
3.3 The style of writing/register conforms to the requirements of the text type and the brief.	0.5
<i>Total</i>	1.5
4. Technical aspects	
4.1 The visual appearance of the text is in line with the brief.	0.5
4.2 The spelling is in line with the target language and the brief.	0.5
4.3 The punctuation is in line with the target language and the brief.	0.5
<i>Total</i>	1.5
5. Content	
5.1 The information in the target text is adequate, given the brief.	1
5.2 The information in the target text is relevant, given the brief.	0.5
<i>Total</i>	1.5
6. Language	
6.1 The use of idiomatic language fulfils the requirements of the target language and the brief.	1
6.2 The use of syntax fulfils the requirements of the target language and the brief	1
6.3 The use of terminology is in line with the requirements of the target language and the brief.	0.5
<i>Total</i>	2.5
7. Grand total	10

Appendix-E. Source Texts Obtained from Daily Star Online Magazine

Text A: Ebola disease could hit ANYWHERE in the world: 'We've been lucky... numbers could EXPLODE'

EBOLA could strike anywhere in the world, as an expert warns we've been lucky so far. [Ebola](#) is currently ravaging parts of Africa, with fears the disease is about to explode in numbers. The outbreak has become the second deadliest in history, infecting approximately three people a day. And health professionals were terrified the disease could spread to Europe after a patient displayed Ebola-like symptoms in Sweden. Although the outbreak was contained, experts still fear the disease could cross continents and cause devastation. They are of the opinion that it could hit anywhere in the world. Harvard health scholar Dr. Ashish Jha told Daily Star Online how easily the disease could be spread.

He told us: "While there's an outbreak happening there's always the chance the somebody could get on a plane and transmit the disease.

"As the disease gets worse in the DRC, the greater the chance is of that happening.

"It's the same for the US, and that goes for China and India too.

"We've been very lucky so far."

So far, 488 people have died in this Ebola epidemic, making it the second worst outbreak ever.

Although Dr. Jha does not believe it will become the deadliest ever, where 11,310 people were killed, but the death toll could still "explode".

He added: "Because of the fighting going on in the area, this outbreak could explode in numbers. "Fighting stops medical personnel from doing their jobs."

"It only takes then disease to spread into another country, then a whole new set of people could get infected."

Numerous militia groups are at war in the north-east of Congo, already ravaged by the disease. Reports have even claimed the groups are attacking health workers trying to treat people.

Text B: A WOMAN is beginning a prison sentence for texting her suicidal boyfriend encouraging him to kill himself.

Michelle Carter, from Massachusetts, US, sent her boyfriend a barrage of text messages urging him to take his own nearly life five years ago. Carter ordered Conrad Roy III to get back in his van after it started filling with carbon monoxide fumes. Carter was only 17 at the time of the death in July 2014. In her texts, she wrote: "You're finally going to be happy in heaven. No more pain.

"It's OK to be scared and it's normal."

Mr. Roy, a former Marine, was suicidal and had previously been admitted to a psychiatric hospital. The 22-year-old was taken into custody on Monday after a top court last week upheld her manslaughter conviction in the 18-year-old's death. She will now begin her 15-month prison sentence. Carter was originally sentenced in 2017 after a judge ruled that she caused Roy's death after telling him to "get back in" his truck that was filling with toxic gas after he told her he was scared. But the judge allowed her to remain free while she appealed her conviction. In a ruling this month, the Massachusetts Supreme Judicial Court shot down Carter's appeal, saying the evidence showed that her conduct caused Roy's death by suicide. The case raised free speech questions about whether a person should be held responsible through their words for someone else's actions. Justice Scott Kafker said in a statement: "The evidence against the defendant proved that, by her wanton or reckless conduct, she caused the victim's death by suicide."

"After she convinced him to get back into the carbon monoxide-filled truck, she did absolutely nothing to help him," Kafker went on.

"She did not call for help or tell him to get out of the truck as she listened to him choke and die."

In court documents, Carter admitted Roy's death was her fault, and that she "could have stopped him".

Text C: Shameless actor dies aged 58 after battle with cancer

SHAMELESS actor Steve Bean has died at the age of fifty-eight. The actor has died from a rare form of nose cancer. The star, who played Abe, passed away on January 21 following a long-running battle with nose cancer. The star, who played Abe, passed away on January 21

following a long-running battle with nose cancer. He was diagnosed with Sino-Nasal Squamous Cell Carcinoma, one of the rarest forms of cancer, which caused him to lose his nose.

Steve spoke about his cancer treatment with *Mel* magazine, revealing that he suffered from severe nosebleeds and congestion in 2016. His wife Caroline Carrigan also told the *Daily Mail*: "The tumor, unbeknownst to us, was already pushing his nose literally out of joint and he was starting to look like a boxer. And then like something from *Avatar*. "I did not expect to be told they were going to remove his nose. You can imagine what that does to a working actor." Steve was given 9 to 12 months to live after chemotherapy and radiation therapy failed to work. He also had his eye removed for his own comfort as the tumor grew and began to press into it. He is survived by his son Jacob Randall Levy. Steve was born Stephen Joseph Levy in Lynn, Massachusetts, in April 1960, and he was raised in Providence, Rhode Island. He started performing while attending Carnegie Mellon University in Pittsburgh, appearing in local nightclubs. The comedian teamed up with Chris Zito in 1980, forming a duo known as Zito & Bean, which gained a large following in the Boston comedy scene. He also appeared in *Ray Donovan* in 2016, and also had small parts in *Veronica Mars*, *Cheers* and *Married with Children*.

Text D: Elon Musk reveals how much a ticket to Mars will cost and people are excited

ELON Musk has revealed how much a ticket to Mars will cost. The billionaire SpaceX CEO Elon Musk is currently in a race against NASA to get humans to the red planet. And when quizzed about how much it will cost to get to Mars, the 47-year-old finally revealed the estimate. Replying to a question on Twitter, Musk said it should hopefully cost less than £500,000. He even suggested that one could hitch a ride on a Mars-bound spacecraft for less than £100,000 before long.

"[The price is] very dependent on volume, but I'm confident moving to Mars (return ticket is free) will one day cost less than \$500k & maybe even below \$100k," Musk explained.

"Low enough that most people in advanced economies could sell their home on Earth & move to Mars if they want." Those hoping to bag a ticket were impressed with the price.

One replied: "Count me in," while another said: "Wow, that is not as much as I would have thought."

A number of Twitter users asked the boss what the public could do to support its development, apparently in the hope of being allowed to make trips to Mars as soon as possible.

"The goodwill of the public is critical to Starship's success," he added.

He was then asked how he would ensure the public stayed engaged with the project.

"Just planning on keeping the public informed about progress & setbacks," he wrote about his plans. "Will be some RUDs along the way, but excitement is guaranteed."

Musk has long promised that living on Mars will be a reality within our lifetimes, even suggesting that he himself might move there at some point in the not-too-distant future.

He hopes that space travelers will eventually make their journey on SpaceX's spaceship, a new craft that is going through the testing process.



Appendix-F. TAP Evaluation Guidelines for Raters

Purpose

Before starting the evaluation process of the translation papers, please read the guidelines below carefully.

These guidelines are provided to help guide you and others in producing think-aloud protocols for this research in a consistent and informative manner. Think-aloud protocols ask people to say everything they are thinking while they perform a task in order to document and better understand what raters pay attention to and consider important when they complete a task. The purpose of the think-aloud protocols for this study is to find out in as much detail as possible what you as an assessor of EFL translation papers are thinking, deciding, and doing while rating a sample of translation papers. The most important thing to remember is to say everything you are thinking, and to make certain this is recorded clearly onto the voice-recorder. What you say will become important data for my research.

Thank you in advance!

The Assessment Task

You will receive a package of 80 translation papers produced by second-year university EFL students with upper-intermediate level language proficiency. However, you will assess 32 of them using think-aloud protocol while the other 48 translation papers will be assessed following standard procedures. The essays that you will assess by thinking aloud have been tagged at the top of the page with the notification 'TAP'. Those essays will also be listed at the end of this document. Say as much as you can while you are reading the translation paper and deciding on how to rate it, and be sure the score you assign to each translation paper is recorded along with your ongoing impressions of the paper.

The Translation Papers

You will also receive copies of the source texts that were originally given to the students so that you can compare their translations with the original texts. There are four source texts with different topics. From each text, there are 20 translation papers in your packs. The translation papers have been identified with code numbers. The order in which you receive the papers has been sequenced randomly in quality, but you should receive compositions of distinct qualities.

The Scorings

While evaluating translation papers, try to use the holistic rubric and the checklist as the basis for your decision. The rating will not be judged as right or wrong. However, I will be analyzing the scores that you assign to the translation papers along with the spoken data regarding your thoughts while assessing the papers.

Voice-recording Your Thoughts

- Keep talking, conveying your thoughts continuously while you assess the translation papers beginning from the moment you first see the paper until you have completed rating it.
- Feel free to speak in either English or Turkish. If you speak in Turkish, it will be translated into English for the final analysis of the data.
- Speak continuously. Report fully, even what might seem trivial. Do not assume that others know what you are doing or thinking.
- Try to avoid speech fillers (i.e., uh, um) as much as possible. Try to use words instead, so that I can understand what your thoughts are.
- Talk and make your assessment as naturally and as honestly as you can, according to what you usually do when you assess students' translation papers. Don't start rationalizing your ideas at length; I am just interested in your natural thought process as you make decisions.

Instructions for Recording

1. Turn on the voice-recorder or your smart phone so that you can record your voice and check that it works. Check whether it records properly and that the quality of the recording is clear by trying out a few words initially, then playing it back. Make sure there is no background noise (e.g., fans, music, foot tapping, etc.).
2. Keep the recorder/smart phone at an appropriate distance from your face and be sure it captures your voice clearly.
3. Turn on the recorder/smart phone, and state the essay code and your name (or pseudonym to be used in the research) at the beginning of each translation paper.
4. While rating the essay, follow the instructions above (*Voice-recording Your Thoughts*). Then, when you have made a rating decision, indicate the score (out of 10) that you have assigned to the translation paper.

5. Report your first impression of the essay and whether it influences your rating. Then continue talking—saying what you think—as you are making your assessment decisions.
6. You may read the essays aloud or silently according to what feels most “natural” to you. Make sure you report exactly what you are doing. If you are reading silently, indicate which part of the paper elicits your comments.
7. If you happen to reconsider any of your ratings (e.g., for a second or third time), verbalize your reason(s) for doing so and indicate on the recorder that this is what you are doing.
8. If you have to take a break while you are assessing the papers, indicate on the recorder that you are doing this, turn the voice recorder off or pause the recording on your smart phone. Then, when you start again, indicate this clearly on the device.
9. When you have completed assessing the translation paper, turn off the voice-recorder/smart phone.
10. Please record your thoughts for each translation paper separately.
11. At the end of the assessment session and voice-recording, please put all the papers together with the recorder that you used (if you were provided with one) back into the package. Thank you!

(adapted from Cumming, Kantor, & Powers, 2001, pp. 83-85)

THINK-ALoud PROTOCOLS WILL BE USED FOR THE FOLLOWING TRANSLATION PAPERS: 6A, 12A, 13A, 14A, 17A, 18A, 19A, 20A, 2B, 5B, 9B, 10B, 11B, 14B, 16B, 19B, 1C, 2C, 4C, 8C, 12C, 15C, 16C, 17C, 3D, 5D, 6D, 9D, 10D, 11D, 18D, 20D

Appendix-G. Official Permission Form Obtained from Kafkas University Faculty of Science and Letters



T.C.
KAFKAS ÜNİVERSİTESİ REKTÖRLÜĞÜ
Öğrenci İşleri Daire Başkanlığı

Sayı : 10829923-10.99-E.7894
Konu : Arş.Gör.Nilüfer AYBİRDİ

14/03/2019

ATATÜRK ÜNİVERSİTESİ REKTÖRLÜĞÜNE
(Öğrenci İşleri Daire Başkanlığı)

İlgi : Öğrenci İşleri Daire Başkanlığının 27.02.2019 tarih ve E.1900066389 sayılı yazısı.

Üniversiteniz Eğitim Bilimleri Enstitüsü Yabancı Diller Eğitimi Anabilim Dalı Doktora öğrencisi Nülüfer AYBİRDİ'nin 01 Mart – 30 Nisan 2019 tarihleri arasında tez çalışması uygulamasını Üniversitemiz Fen-Edebiyat Fakültesi Batı Dilleri ve Edebiyatları Bölümünde yapması Rektörlüğümüzce uygun görülmüştür.

Gereğini bilgilerinize arz ederim.

e-imzalıdır

Prof.Dr.Engin KILIÇ
Rektör a.
Rektör Yardımcısı

Ek : (1 Sayfa)



Kafkas Üniversitesi Rektörlüğü Turan Çelik Caddesi Merkez/KARS
Telefon: (474) 225 11 50 Faks: (474) 225 11 61 E-Posta: info@kafkas.edu.tr
www.kafkas.edu.tr

Ayrıntılı Bilgi İçin: Gülseren ÇAKAR
Tel: E-Posta:

Bu belge 5070 sayılı e-İmza Kanununa göre Prof.Dr.Engin KILIÇ tarafından 14.03.2019 tarihinde e-imzalanmıştır.
Evrağınızı belgesorgu.kafkas.edu.tr linkinden 60D12C89XE kodu ile doğrulayabilirsiniz.

T.C.
KAFKAS ÜNİVERSİTESİ
FEN EDEBİYAT FAKÜLTESİ
Batı Dilleri ve Edebiyatları Bölüm Başkanlığı

SAYI: 12046136-903- 10

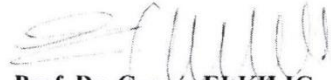
KONU: Uygulama İzni (Nilüfer AYBİRDİ)

07.03.2019

DEKANLIK MAKAMINA

Bölümümüz öğretim elamanı Arş. Gör. Nilüfer AYBİRDİ, tez çalışmasının uygulamasını Bölümümüzde yapması uygundur.

Gereğini bilgilerinize arz ederim.


Prof. Dr. Gençer ELKILIÇ

Batı Dilleri ve Edebiyatları Bölüm Başkanı

Ekler:

6 (Sayfa)

T.C.
KAFKAS ÜNİVERSİTESİ
FEN EDEBİYAT FAKÜLTESİ
DEKANLIK MAKAMINA
Kayıt Tarihi: 7.3.2019
Kayıt No: 4822

Personel
08.03.2019
[Signature]

Bu belge 5070 sayılı e-İmza Kanununa göre Prof.Dr.Engin KILIC tarafından 14.03.2019 tarihinde e-imzalanmıştır. Evrağınızı belgesorgu.kafkas.edu.tr linkinden 60D12C89XE kodu ile doğrulayabilirsiniz.



T.C.
ATATÜRK ÜNİVERSİTESİ REKTÖRLÜĞÜ
Öğrenci İşleri Daire Başkanlığı



Sayı : 88179374-300-E.1900066389
Konu : Uygulama İzni (Nülüfer
AYBİRDİ)

27.02.2019

KAFKAS ÜNİVERSİTESİ REKTÖRLÜĞÜNE

Üniversitemiz Eğitim Bilimleri Enstitüsü Yabancı Diller Eğitimi Ana Bilim Dalı İngiliz Dili Eğitimi Bilim Dalı doktora öğrencisi Nülüfer AYBİRDİ'nin tez çalışmasının uygulamasını 01.03.2019-30.04.2019 tarihleri arasında Üniversitenizde yapma talebi ile ilgili adı geçen Enstitünün 26.02.2019 tarih ve 56785782-302.08.01-E.1900065461 sayılı yazısı ekte sunulmuştur.

Bilgilerinizi ve adı geçenin tez çalışması için gerekli iznin verilip verilemeyeceğinin tarafımıza bildirilmesi hususunda gereğini arz ederim.

Prof.Dr. Ayhan ÇELİK
Rektör V.

Ek : 26.2.2019 tarihli 56785782-302.08.01-E.1900065461 sayılı belge

Atatürk Üniversitesi Merkez Yerleşkesi 25240 Erzurum
Tel: +90 442 2311601
Elektronik Ağ: <http://www.atauni.edu.tr/#?birim=ogrenci-isleri-daire-baskanligi>
Kep Adresi: atauni@hs01.kep.tr

Bilgi: Cevahir KARTA
Faks: +90 442 2361026
E-Posta: odaire@atauni.edu.tr



Bu belge, 5070 sayılı Elektronik İmza Kanununun 5. maddesi gereğince güvenli elektronik imza ile imzalanmıştır.
<https://ubys.atauni.edu.tr/ERMS/Record/Confirmation/Confirmation?code=691EFD0>



T.C.
ATATÜRK ÜNİVERSİTESİ REKTÖRLÜĞÜ
Eğitim Bilimleri Enstitüsü Müdürlüğü



Sayı : 56785782-302.08.01-E.1900065461

26.02.2019

Konu : Uygulama İzni (N. AYBİRDİ)

REKTÖRLÜK MAKAMINA
(Öğrenci İşleri Daire Başkanlığı)

İlgi : 20.02.2019 tarihli ve 29202147-000-E.1900059200 sayılı belge.

Danışmanlığını Prof Dr. Mehmet TAKKAÇ'ın yürüttüğü Yabancı Diller Eğitimi Ana Bilim Dalı İngiliz Dili Eğitimi Bilim Dalı doktora öğrencisi **Nülüfer AYBİRDİ**'nin tez çalışması kapsamında 01.03.2019-30.04.2019 tarihleri arasında uygulama yapabilmesi için Kafkas Üniversitesi Fen Edebiyat Fakültesinden gerekli izinlerin alınması hususunda gereğini arz ederim.

Prof.Dr. Mustafa SÖZBİLİR
Müdür

Ek : Dilekçe ve Ekleri

Atatürk Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğü 25240 Erzurum
Tel: +90 442 2315630
Elektronik Ağ: <http://www.atauni.edu.tr/#?birim=egitim-bilimleri-enstitusu>

Kep Adresi: atauni@hs01.kep.tr

Bilgi: Nilgün ŞAHİN
Faks: +90 442 2315883
E-Posta: egitbil@atauni.edu.tr



Bu belge, 5070 sayılı Elektronik İmza Kanununun 5. maddesi gereğince güvenli elektronik imza ile imzalanmıştır.
<https://ubys.atauni.edu.tr/ERMS/Record/Confirmation/Confirmation?code=092BE17>

ATATÜRK ÜNİVERSİTESİ

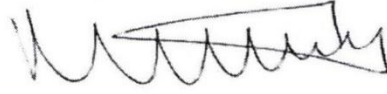
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI BAŞKANLIĞINA

Danışmanlığımı yürüttüğüm 16030806001 numaralı doktora öğrencisi Nilüfer AYBİRDİ'nin "Üniversite Öğrencilerinin Çeviri Kalitesinin Değerlendirilmesi Sürecinin İncelenmesi" başlıklı tez çalışmasının uygulamasının 01.03.2019-30.04.2019 tarihleri arasında yapılması için Kafkas Üniversitesi Fen Edebiyat Fakültesi İngiliz Dili ve Edebiyatı Bölümünden gerekli izinlerin alınması hususunda;

Gereğini bilgilerinize arz ederim.

19.02.2019

Prof. Dr. Mehmet TAKKAÇ



EK 1: Uygulamada kullanılacak 4 adet metin

Woman, 22, starts jail time for texting suicidal boyfriend encouraging him to kill himself

A WOMAN is beginning a prison sentence for texting her suicidal boyfriend encouraging him to kill himself.

Michelle Carter, from Massachusetts, US, sent her boyfriend a barrage of text messages urging him to take his own nearly life five years ago. Carter ordered Conrad Roy III to get back in his van after it started filling with carbon monoxide fumes. Carter was only 17 at the time of the death in July 2014. In her texts, she wrote: "You're finally going to be happy in heaven. No more pain."

"It's OK to be scared and it's normal."

Mr Roy, a former Marine, was suicidal and had previously been admitted to a psychiatric hospital. The 22-year-old was taken into custody on Monday after a top court last week upheld her manslaughter conviction in the 18-year-old's death. She will now begin her 15-month prison sentence. Carter was originally sentenced in 2017 after a judge ruled that she caused Roy's death after telling him to "get back in" his truck that was filling with toxic gas after he told her he was scared. But the judge allowed her to remain free while she appealed her conviction. In a ruling this month, the Massachusetts Supreme Judicial Court shot down Carter's appeal, saying the evidence showed that her conduct caused Roy's death by suicide. The case raised free speech questions about whether a person should be held responsible through their words for someone else's actions. Justice Scott Kafker said in a statement: "The evidence against the defendant proved that, by her wanton or reckless conduct, she caused the victim's death by suicide."

"After she convinced him to get back into the carbon monoxide-filled truck, she did absolutely nothing to help him," Kafker went on.

"She did not call for help or tell him to get out of the truck as she listened to him choke and die."

In court documents, Carter admitted Roy's death was her fault, and that she "could have stopped him".

Passage statistics: Number of sentences, 18; Words per sentence, 18.28; Percentage of difficult words, 9.12%; Flesch reading ease, 66.4; Flesch-Kincaid grade level, 8.54.

Elon Musk reveals how much a ticket to Mars will cost and people are excited

ELON Musk has revealed how much a ticket to Mars will cost. The billionaire SpaceX CEO Elon Musk is currently in a race against NASA to get humans to the red planet. And when quizzed about how much it will cost to get to Mars, the 47-year-old finally revealed the estimate. Replying to a question on Twitter, Musk said it should hopefully cost less than £500,000. He even suggested that one could hitch a ride on a Mars-bound spacecraft for less than £100,000 before long.

"[The price is] very dependent on volume, but I'm confident moving to Mars (return ticket is free) will one day cost less than \$500k & maybe even below \$100k," Musk explained.

"Low enough that most people in advanced economies could sell their home on Earth & move to Mars if they want." Those hoping to bag a ticket were impressed with the price.

One replied: "Count me in," while another said: "Wow, that is not as much as I would have thought."

A number of Twitter users asked the boss what the public could do to support its development, apparently in the hope of being allowed to make trips to Mars as soon as possible.

"The goodwill of the public is critical to Starship's success," he added.

He was then asked how he would ensure the public stayed engaged with the project.

"Just planning on keeping the public informed about progress & setbacks," he wrote about his plans. "Will be some RUDs along the way, but excitement is guaranteed."

Musk has long promised that living on Mars will be a reality within our lifetimes, even suggesting that he himself might move there at some point in the not-too-distant future.

He hopes that space travellers will eventually make their journey on SpaceX's spaceship, a new craft that is going through the testing process.

Passage statistics: Number of sentences, 17; Words per sentence, 18; Percentage of difficult words, 8.17%; Flesch reading ease, 69.96; Flesch-Kincaid grade level, 8.

Appendix-H. Written Consent Form (for raters)

Statement of Consent

By signing this form, it proves that:

- You have received explanation from the researcher regarding the background, purpose, significance, and procedures of this research;
- You have understood the details of your participation in this research;
- You are aware of the benefits and risks that you will experience;
- You have been given opportunities to ask questions, and the researcher has responded to them satisfactorily;
- You are aware that you are free to withdraw your participation from this research
- You are aware that the scoring sheets that you will produce will be included in the researcher's thesis, and you agree to this;
- You are aware that the TAPs will be transcribed, and included in the researcher's thesis, and you agree to this;
- You understand that the data you provide for this thesis may be used by other people in the future as comparison for their own research;
- You confirm that you are willingly and voluntarily participate in this research with no coercion or undue influences from any parties.

Name: _____

Signature: _____

Appendix-I. (p x r x q) Random Effects G-study for All Translation Papers

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PAPER	P	40	INF	
RATER	R	10	INF	
QUALITY	Q	2	INF	

Analysis of Variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	3935.78075	39	100.91746	4.84629	4.84629	4.84629	61.8	1.11473
R	128.18225	9	14.24247	-0.33339	-0.33339	-0.33339	0.0	0.22101
Q	24.29045	1	24.29045	-0.03928	-0.03928	-0.03928	0.0	0.06467
PR	1027.53575	351	2.92745	0.98762	0.98762	0.98762	12.6	0.11586
PQ	78.63755	39	2.01635	0.10641	0.10641	0.10641	1.4	0.04511
RQ	350.44755	9	38.93862	0.94966	0.94966	0.94966	12.1	0.41509
PRQ	334.22445	351	0.95221	0.95221	0.95221	0.95221	12.1	0.07167
Total	5879.09875	799					100%	

Measurement Design P/RQ

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	4.84629					
		R			(0.00000)	0.0
		Q			(0.00000)	0.0
		PR	0.09876	49.5	0.09876	40.0
		PQ	0.05321	26.7	0.05321	21.5
		RQ			0.04748	19.2
		PRQ	0.04761	23.9	0.04761	19.3
Sum of variances	4.84629		0.19958	100%	0.24706	100%
Standard deviation	2.20143		Relative SE: 0.44674		Absolute SE: 0.49705	
Coef_G relative	0.96					
Coef_G absolute	0.95					

Appendix-J. (p x r) Random Effects G-study for High-Quality Translation Papers

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PAPER	P	40	INF	
RATER	R	10	INF	

Analysis of Variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	183.22440	39	4.69806	0.38612	0.38612	0.38612	13.1	0.10395
R	626.95190	9	69.66132	1.72061	1.72061	1.72061	58.5	0.74259
PR	293.72810	351	0.83683	0.83683	0.83683	0.83683	28.4	0.06299
Total	1103.90440	399					100%	

Measurement Design P/R

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.38612					
		R			0.17206	67.3
		PR	0.08368	100.0	0.08368	32.7
Sum of variances	0.38612		0.08368	100%	0.25574	100%
Standard deviation	0.62139		Relative SE: 0.28928		Absolute SE: 0.50571	
Coef_G relative	0.82					
Coef_G absolute	0.60					

Appendix-K. (p x r) Random Effects G-study for Low-Quality Translation Papers

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PAPER	P	40	INF	
RATER	R	10	INF	

Analysis of Variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	104.46990	39	2.67872	0.17205	0.17205	0.17205	6.2	0.05960
R	595.87690	9	66.20854	1.63126	1.63126	1.63126	59.1	0.70579
PR	336.31710	351	0.95817	0.95817	0.95817	0.95817	34.7	0.07212
Total	1036.66390	399					100%	

Measurement Design P/R

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.17205					
		R			0.16313	63.0
		PR	0.09582	100.0	0.09582	37.0
Sum of variances	0.17205		0.09582	100%	0.25894	100%
Standard deviation	0.41479		Relative SE: 0.30954		Absolute SE: 0.50886	
Coef_G relative	0.64					
Coef_G absolute	0.40					

Appendix-L. Reliability and Dependability Coefficients of Low- and High-Quality Translation Papers by Gender

High-Quality Papers: Measurement Design (P/R) for Female Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.46529					
		R			0.74942	80.7
		PR	0.17912	100.0	0.17912	19.3
Sum of variances Standard deviation	0.46529		0.17912	100%	0.92854	100%
	0.68212		Relative SE: 0.42323		Absolute SE: 0.96361	
Coef_G relative	0.72					
Coef_G absolute	0.33					

High-Quality Papers: Measurement Design (P/R) for Male Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.31785					
		R			0.20012	56.3
		PR	0.15543	100.0	0.15543	43.7
Sum of variances Standard deviation	0.31785		0.15543	100%	0.35554	100%
	0.56378		Relative SE: 0.39424		Absolute SE: 0.59628	
Coef_G relative	0.67					
Coef_G absolute	0.47					

Low-Quality Papers: Measurement Design (P/R) for Female Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.23503					
		R			0.20861	45.3
		PR	0.25195	100.0	0.25195	54.7
Sum of variances Standard deviation	0.23503		0.25195	100%	0.46056	100%
	0.48479		Relative SE: 0.50195		Absolute SE: 0.67864	
Coef_G relative	0.48					
Coef_G absolute	0.34					

Low-Quality Papers: Measurement Design (P/R) for Male Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.13398					
		R			0.33951	68.9
		PR	0.15353	100.0	0.15353	31.1
Sum of variances	0.13398		0.15353	100%	0.49304	100%
Standard deviation	0.36603		Relative SE: 0.39183		Absolute SE: 0.70217	
Coef_G relative	0.47					
Coef_G absolute	0.21					



Appendix-M. Reliability and Dependability Coefficients of Low- and High-Quality Translation Papers by Experience

High-Quality Papers: Measurement Design (P/R) for Experienced Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.14128					
		R			0.25126	60.2
		PR	0.16578	100.0	0.16578	39.8
Sum of variances Standard deviation	0.14128		0.16578	100%	0.41704	100%
	0.37588		Relative SE: 0.40716		Absolute SE: 0.64579	
Coef_G relative	0.46					
Coef_G absolute	0.25					

High-Quality Papers: Measurement Design (P/R) for Inexperienced Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.65987					
		R			0.44718	73.3
		PR	0.16317	100.0	0.16317	26.7
Sum of variances Standard deviation	0.65987		0.16317	100%	0.61035	100%
	0.81232		Relative SE: 0.40394		Absolute SE: 0.78125	
Coef_G relative	0.80					
Coef_G absolute	0.52					

Low-Quality Papers: Measurement Design (P/R) for Experienced Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.03387					
		R			0.42455	65.7
		PR	0.22123	100.0	0.22123	34.3
Sum of variances Standard deviation	0.03387		0.22123	100%	0.64578	100%
	0.18405		Relative SE: 0.47035		Absolute SE: 0.80360	
Coef_G relative	0.13					
Coef_G absolute	0.05					

Low-Quality Papers: Measurement Design (P/R) for Inexperienced Raters

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.22213					
		R			0.30075	62.6
		PR	0.17966	100.0	0.17966	37.4
Sum of variances	0.22213		0.17966	100%	0.48041	100%
Standard deviation	0.47131		Relative SE: 0.42386		Absolute SE: 0.69312	
Coef_G relative	0.55					
Coef_G absolute	0.32					

CURRICULUM VITAE

Personal Information

Name & Surname: Nilüfer AYBİRDİ

Place of Birth: Selim/Kars

Date of Birth: 10.06.1987

Educational Background

Bachelor's Degree: Department of English Language Teaching, Education Faculty, Gazi University

Master's Degree: Department of English Language and Literature, Graduate School of Social Sciences, Kafkas University

Doctorate Degree: Department of English Language Teaching, Graduate School of Educational Sciences, Atatürk University

Publications

a) International Books and Book Chapters

Han, T., & **Aybirdi, N.** (2017). Turkish EFL students' perspectives regarding responsibilities in reducing foreign language classroom anxiety. In D. Köksal (Ed.), *Researching ELT: Classroom methodology and beyond* (pp. 115-126). Frankfurt: Peter Lang.

Han, T., & **Aybirdi, N.** (2016). 21st century skills for EFL learners. In İ. Yaman, E. Ekmekçi & M. Şenel (Eds.), *Current trends in ELT* (pp. 42-57). Ankara: Nüans Publishing.

Han, T., & **Aybirdi, N.** (2015). A Literature Review on Teaching Turkish as a Foreign Language. In T. Han (Ed.), *Interdisciplinary Humanities Research: A Turkish Perspective* (pp. 42-57). NY: Untested Ideas Research Center.

b) International Refereed Research Journals

Barın, M., Han, T., & **Aybirdi, N.** (2018). İngilizceyi yabancı dil olarak öğrenen Türk öğrencilerin etik olmayan davranışları üzerine bir araştırma. *Kastamonu Eğitim Dergisi*, 26(5), 1785-1792.

- Elkılıç, G., & **Aybirdi, N.** (2016). Metaphorical perceptions of Turkish EFL instructors in relation to their students. *The European Proceedings of Social & Behavioural Sciences epsbs*, 330-337.
- Kaya, H. I., Han, T., & **Aybirdi, N.** (2015). Prospective EFL teachers' attitudes towards the use of authentic materials in EFL reading classes. *Journal of International Social Research*, 8(37), 773-783.
- Han, T., Takkaç-Tulgar, A., & **Aybirdi, N.** (2019). Factors causing demotivation in EFL learning process and the strategies used by Turkish EFL learners to overcome their demotivation. *Advances in Language and Literary Studies*, 10(2), 56-65.
- Han, T., Elkılıç, G., & **Aybirdi, N.** (2019). Eşlerin birbirlerine sosyo-edimsel hitap biçimleri üzerine tanımlayıcı bir çalışma. *Milli Folklor*, 16(123), 166-176.

Work Experience

2012-2013: EFL Instructor, Ardahan University

2014-Present: Research Assistant, Kafkas University, Department of English Language and Literature

Contact

E-mail Address: niluferaybirdi@hotmail.com